DEBRIS, RUBBLE PILES AND FAÇADE DAMAGE DETECTION USING MULTI-RESOLUTION OPTICAL REMOTE SENSING IMAGERY

Diogo André Vicente Amorim Duarte

DEBRIS, RUBBLE PILES AND FAÇADE DAMAGE DETECTION USING MULTI-RESOLUTION OPTICAL REMOTE SENSING IMAGERY

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on 23rd January 2020 at 12:45 hrs

by

Diogo André Vicente Amorim Duarte
born on 27th December 1987
in Pombal, Portugal

This thesis has been approved by
Prof.dr.ir. M.G. Vosselman
Prof.dr. N. Kerle
Dr.ir. F.C. Nex

ITC dissertation number 374
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4940-0
DOI 10.3990/1.9789036549400

Cover designed by
Printed by ITC Printing Department

Graduation committee:

Chairman/Secretary

Prof.dr.ir. A. Veldkamp	University of Twente

Supervisors

Prof.dr.ir. M.G. Vosselman	University of Twente / ITC
Prof.dr. N. Kerle	University of Twente / ITC

Co-supervisor

Dr.ir. F.C. Nex	University of Twente / ITC

Members

Prof.dr. A.K. Skidmore	University of Twente / ITC
Prof.dr. M.K. van Aalst	University of Twente / ITC
Prof.dr. P. Gamba	University of Pavia, Italy
Prof.dr. S. Lefevre	University of South Brittany, France

To my mother São, my father Nabeto and my sister Barbara

Summary

Knowledge of the location of damaged buildings is of utmost importance for both the response and recovery phases of the disaster management cycle. To this regard, remote sensing images have been continuously used over the last 20 years as the main data source in approaches to detect building damages. Partially and totally collapsed buildings are the structures which might contain entrapped victims; hence many studies focus on the mapping of debris and rubble piles. Nonetheless such assumption might leave out damage evidences such as spalling or cracks, especially in the façades. When comparing with the mapping of rubble piles and debris, the façade damage detection is an understudied topic. The objective of the research reported in this thesis was focused on the mapping of both debris/ruble piles and façade damages from remote sensing images.

The mapping of partially and totally collapsed buildings is often constrained by the used system (platform and sensor). There is a growing amount of imagery being collected (e.g. by the International Charter and Emergency Management Service) using different sensors, platforms and resolution, where their optimal use and integration would represent an opportunity to positively impact the detection of building damages. However, this multitude of systems does not imply the availability of large datasets sufficient to train recent and more complex algorithms such as convolutional neural networks (CNN). Hence, one of the goals of this thesis is to fuse satellite and aerial (manned and unmanned) image samples in a unique classification network to assess the building damage detection in each of the considered resolution levels.

While there are several contributions regarding the mapping of debris and rubble piles, this is not the case when focusing on the specific case of façade damages. Nonetheless, façade image data are already being collected by both aerial manned and unmanned vehicles. Regarding the use of UAV, only a few approaches focused on the specific issue of façade damage detection. These are often not made operational and require computationally expensive procedures which limit their utility to stakeholders, who need fast and reliable façade damage information. One of the objectives of this thesis is to improve the efficiency of such façade damage detection procedures. On the other hand, aerial manned platforms have a wider coverage whilst capturing data at a lower resolution. In particular, the use of imagery coming from aerial (manned) oblique surveys has substantially increased in the last decade, leading to periodic aerial surveys over entire cities in many countries. Such data could be therefore exploited for multi-temporal image classification of façade damages over a given city/region. This was the main focus of the third goal of this thesis, the detection of façade damages mainly focusing on the use of multi-temporal aerial oblique imagery to infer on the damage state of a given façade.

Related with the overall objective of mapping rubble piles, debris and façade damages, three distinct objectives, with their own set of experiments are investigated in this thesis:

1. **Mapping of partially and totally collapsed buildings using multi-resolution remote sensing images (Chapters 2 and 3)**.
   A preliminary study regarding the use of multi-resolution imagery focused on the specific case of the satellite image classification of building damages. Features were extracted from satellite and aerial (manned and unmanned) imagery and fed to a supervised classifier to detect rubble piles and debris in satellite images. The approaches considering image samples coming from other resolutions outperformed the traditional approach by nearly 4%, where the traditional approaches used only satellite image samples during training. Picking up on these results, the approach was extended to the other resolutions, referring to aerial manned and unmanned. Using the multi-resolution approach for the image classification of debris and rubble piles, improved the results in the case of aerial unmanned (by ~5 %) and performed similarly to traditional approaches when using aerial manned platforms. The best performing multi-resolution approach merged the features coming from the three different sets of images, and also considered feature information from the intermediate layers of each of the levels of resolution. The approach was also tested for geographical transferability where the differences between the traditional and multi-resolution approaches were maintained.

2. **Efficient detection of earthquake induced façade damages from UAV images (Chapter 4):**
   An approach to perform a more efficient detection of façade damages from UAV images was developed. It aimed at reducing the time between the deployment of the UAV and per façade damage results, in order to be of use to first responders Such efficiency was achieved by directing all damage classification computations to the specific image regions containing the façades. This was achieved by acquiring nadir images in a first flight, which allowed to detect the buildings and consequently define the façades. This 3D façade information was then used to identify the façades in oblique images acquired in a second flight from which façade damages could be assessed. The buildings were identified by segmenting the building roofs from the sparse point cloud directly, avoiding the computationally expensive dense image matching algorithm. The acquired data were georeferenced using the on-board information. The second flight was performed only on façades of interest, where all the damage detection procedures were only applied to these same façades. Although this method is more efficient, the detection of façade damages used a model trained only on rubble piles and debris, delivering a high rate of false positives and

leaving out smaller cues of damage such as spalling or cracks. This method is only achieving ~80% accuracy.

3. **Multi-temporal façade damage detection (Chapters 5 and 6)**:
   The last objective focused on the use of multi-temporal aerial oblique datasets to assess a given city/region for façade damages. The first step in the detection of façade damages was the extraction of oblique image patches depicting the façade. To achieve this, the pre-event point cloud was generated through dense image matching, where the rest of the approach followed a similar façade extraction procedure as indicated in 2). Preliminary results on the multi-temporal façade damage detection were obtained by comparing rectified façade image patches, between and within epochs, using a simple cross correlation coefficient. This multi-temporal study was further investigated by integrating it in a supervised classification approach using CNN. This approach focused on two main issues: (i) the optimal fusion the multi-temporal data and (ii) the use of high overlapping aerial images to extract the same façade from different views and embedding them in the multi-temporal approach. The results demonstrated the benefits given by façade damage detection approaches using multi-temporal datasets. Moreover, the results show that considering several views per façade within a CNN approach improves the image classification of façade damages. The multi-temporal approach outperformed the mono-temporal ones by 20% in f1-score, where the best multi-temporal approach achieved an f1-score of 82%. Given the limited number of samples and the relatively low resolution, smaller damage evidences such as small cracks and/or small areas of spalling could not be detected.

The research reported in this thesis was part of the EU (7th Framework Programme) funded INACHUS (Technological and Methodological Solutions for Integrated Wide Area Situation Awareness and Survivor Localization to Support Search and Rescue Teams) project (www.inachus.eu). This project aimed at a time reduction of the response phase performed by FR, namely in the identification of entrapped victims after a disaster. The work reported in this thesis focused on the use of aerial imagery to localize damaged buildings over a given region/building block.

Samenvatting

Kennis van de locatie van beschadigde gebouwen is van het grootste belang voor zowel de respons- als de herstelfase van de rampenbestrijdingscyclus. In dit verband zijn de laatste 20 jaar voortdurend satellietbeelden gebruikt als de belangrijkste gegevensbron bij het opsporen van schade aan gebouwen. Gedeeltelijk en volledig ingestorte gebouwen kunnen ingesloten slachtoffers bevatten; vandaar dat veel studies zich richten op het in kaart brengen van puin en puinhopen. Andere schadebewijzen zoals versplintering of scheuren, met name in de gevels, worden buiten beschouwing gelaten. In vergelijking met het in kaart brengen van puinhopen en puin is de gevelschadedetectie een onderbelicht onderwerp. Het doel van het onderzoek dat in dit proefschrift wordt gerapporteerd was het in kaart brengen van zowel puin/afvalstapels als gevelschade door remote sensing beelden.

Het in kaart brengen van gedeeltelijk en volledig ingestorte gebouwen wordt vaak beperkt door het gebruikte systeem (platform en sensor). Er wordt steeds meer beeldmateriaal verzameld (bijvoorbeeld door de International Charter and Emergency Management Service) met behulp van verschillende sensoren, platforms en resoluties, waarbij het optimale gebruik en de integratie ervan een kans zou bieden om de detectie van schade aan gebouwen te verbeteren. Deze veelheid aan systemen impliceert echter niet dat er grote datasets beschikbaar zijn die voldoende zijn om recente en meer complexe algoritmen zoals convolutionele neurale netwerken (CNN) te trainen. Een van de doelstellingen van dit proefschrift is dan ook om satelliet- en luchtfoto’s (van bemande en onbemande vliegtuigen) samen te voegen in een uniek classificatienetwerk om de detectie van gebouwschade in elk van de beschouwde resolutieniveaus te beoordelen.

Hoewel er verschillende bijdragen zijn met betrekking tot het in kaart brengen van puin en puinhopen, is dit niet het geval voor het specifieke geval van gevelschade. Toch worden de beeldgegevens van gevels al verzameld door zowel bemande als onbemande vliegtuigen. Wat het gebruik van UAVs betreft, waren slechts enkele benaderingen specifiek gericht op het opsporen van gevelschade. Deze worden vaak niet operationeel gemaakt en vereisen rekenkundig dure procedures die het nut ervan beperken voor de belanghebbenden, die behoefte hebben aan snelle en betrouwbare informatie over gevelschade. Een van de doelstellingen van dit proefschrift is het verbeteren van de efficiëntie van dergelijke procedures voor het opsporen van gevelschade. Aan de andere kant hebben de bemande vliegtuigen een groter bereik, terwijl ze gegevens met een lagere resolutie vastleggen. Met name het gebruik van beeldmateriaal dat afkomstig is van oblieke luchtfoto's is de afgelopen tien jaar aanzienlijk toegenomen, wat heeft geleid tot periodieke opname van deze luchtfoto's over hele steden in veel landen. Dergelijke

gegevens zouden dus kunnen worden gebruikt voor een multi-temporele classificatie van gevelschade over een bepaalde stad/regio. Dit was de belangrijkste focus van het derde doel van dit proefschrift, het opsporen van gevelschade, voornamelijk gericht op het gebruik van multi-temporele luchtfoto's om de schade aan een bepaalde gevel af te leiden.

Met betrekking tot de algemene doelstelling van het in kaart brengen van puinhopen, puin en gevelschade worden in dit proefschrift drie verschillende doelstellingen ieder met hun eigen set van experimenten onderzocht:

1. **Het in kaart brengen van gedeeltelijk en volledig ingestorte gebouwen met behulp van multi-resolutie remote sensing beelden (hoofdstuk 2 en 3)**.
   Een voorstudie over het gebruik van multi-resolutiebeelden richtte zich op het specifieke geval van de satellietbeeldclassificatie van schade aan gebouwen. Kenmerken werden uit satelliet- en luchtbeelden (bemand en onbemand) geëxtraheerd en gebruikt in een gecontroleerde classificatie om puinhopen en puin in satellietbeelden op te sporen. De benaderingen waarbij gebruik wordt gemaakt van kenmerken, die afkomstig zijn van andere beeldresoluties, presteerden bijna 4% beter dan de traditionele aanpak, waarbij bij de traditionele benaderingen tijdens de training alleen gebruik werd gemaakt van kenmerken uit satellietbeelden. De aanpak werd uitgebreid naar de andere beelden met resoluties, die met zowel bemand als onbemand vliegtuigen zijn opgenomen. Het gebruik van de multi-resolutiebenadering voor de beeldclassificatie van puin- en puinhopen, verbetert het resultaat in het geval van beelden van onbemande luchtvaartuigen (met ~5%) en blijft ongeveer hetzelfde als bij een traditionele aanpak met luchtfoto's uit bemande vliegtuigen. Bij de best presterende multiresolutiebenadering werden de kenmerken van de drie verschillende reeksen beelden samengevoegd en werd ook rekening gehouden met de informatie over kenmerken van de tussenliggende lagen van elk van de resolutieniveaus. De aanpak werd ook getest op overdraagbaarheid naar andere geografische gebieden, waarbij de verschillen tussen de traditionele en de multi-resolutiebenadering werden gehandhaafd.

2. **Efficiënte detectie van door aardbevingen veroorzaakte gevelschade aan de hand van UAV-beelden (hoofdstuk 4):**
   Er is een aanpak ontwikkeld om een efficiëntere detectie van gevelschade door middel van UAV-beelden uit te voeren. Het doel was om de tijd tussen de inzet van UAVs en de beschikbaarheid van de resultaten over gevelschade te verkorten, zodat de hulpverleners er meer baat bij hebben. Een dergelijke efficiëntie werd bereikt door alle berekeningen van schadeclassificatie te concentreren op de specifieke beelduitsneden die

gevels. Dit werd bereikt door nadirbeelden in een eerste vlucht op te nemen, die het mogelijk maken de gebouwen te detecteren en zo de locaties van gevels te bepalen. De 3D-gevelinformatie werd vervolgens gebruikt om de gevels te identificeren in een oblieke beelden die in een tweede vlucht werden opgenomen en waarin de gevelschade kon worden beoordeeld. De gebouwen werden geïdentificeerd door de daken van de gebouwen rechtstreeks te segmenteren in een ijle puntwolk, waardoor het rekenkundig dure algoritme voor de zgn. dense matching werd vermeden. De verkregen gegevens werden gegeorefereerd aan de hand van de vluchtinformatie. De tweede vlucht werd alleen uitgevoerd op gevels van belang, waarbij alle schadedetectieprocedures alleen op deze gevels werden toegepast. Hoewel deze methode efficiënter is, werd voor het opsporen van gevelschade gebruik gemaakt van een model dat alleen getraind is op puinhopen en puin, waardoor een hoge mate van onjuiste detecties wordt verkregen en kleinere aanwijzingen voor schade zoals versplintering of scheuren worden genegeerd. Deze methode haalde slechts een nauwkeurigheid van ~80%.

3. **Multi-temporele geveldetectie (hoofdstuk 5 en 6):**
   De laatste doelstelling richtte zich op het gebruik van multi-temporele foto's die met bemande vliegtuigen zijn opgenomen, zodat een gehele stad/regio kan worden beoordeeld op gevelschade. De eerste stap in het opsporen van gevelschade was de extractie van uitsneden uit de oblieke foto's die de gevel afbeeldden. Om dit te bereiken werd een puntwolk gegenereerd met dense matching in beelden die voor de aardbeving zijn opgenomen. De rest van de aanpak was vergelijkbaar met de procedure voor gevelextractie zoals aangegeven in 2). Eerste resultaten op de multi-temporele gevelschadedetectie werden verkregen door het vergelijken van gerectificeerde gevelbeelduitsneden van het zelfde en het andere tijdstip met behulp van een eenvoudige kruiscorrelatiecoëfficiënt. Deze multi-temporele studie werd verder onderzocht door het te integreren in een gecontroleerde classificatie met behulp van een CNN. Deze aanpak was gericht op twee belangrijke aspecten: (i) de optimale fusie van de multi-temporele gegevens en (ii) het gebruik van sterk overlappende luchtbeelden om dezelfde gevel uit verschillende aanzichten te halen en in te bedden in de multi-temporale aanpak. De resultaten toonden de voordelen aan van een aanpak voor de detectie van gevelschade met behulp van multi-temporele datasets. Bovendien blijkt uit de resultaten dat het gebruik van meerdere aanzichten per gevel binnen een CNN-aanpak de beeldclassificatie van gevelschades verbetert. De multi-temporele aanpak presteerde 20% beter dan de mono-temporele aanpak in f1-score, waar de beste multi-temporele aanpak een f1-score van 82% haalde. Gezien het beperkte aantal beelduitsneden en de relatief lage resolutie

konden kleinere beschadigingen zoals kleine scheurtjes en/of kleine gebieden met versplintering niet worden opgespoord.

Het onderzoek dat in dit proefschrift wordt gerapporteerd maakte deel uit van het door de EU (7de Kaderprogramma) gefinancierde INACHUS-project (Technological and Methodological Solutions for Integrated Wide Area Situation Awareness and Survivor Localization to Support Search and Rescue Teams, www.inachus.eu). Dit project was gericht op een tijdsvermindering van de responsfase voor de eerste hulpverleners, namelijk bij de identificatie van ingesloten slachtoffers na een ramp. Het werk dat in dit proefschrift wordt gerapporteerd richtte zich op het gebruik van luchtfoto's om beschadigde gebouwen te lokaliseren in een bepaalde regio/woningblok.

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in Enschede and ITC. More specifically I would like to thank my promotor, supervisors, ITC colleagues and staff, and graduation committee members. Would also like to thank the several friends made over these 4 years in Enschede.

I would like to thank Dr. Francesco Nex for his dedication in helping me as he would always manage to arrange a time slot to discuss whatever issue with me. Also, to thank Prof. dr. Norman Kerle for the patience in transmitting knowledge regarding skills and competences within academia. Always with a sharp view when commenting drafts which were the grounds for me to build scientific skills. I would also like to acknowledge Prof. dr. George Vosselman for his support regarding my research focus and scientific contributions.

I would also like to thank all the colleagues at EOS department. The different cultural and scientific backgrounds, of staff and students, made it a very rich experience not only on the academic but also on the personal side.

Grateful to have Rita's support throughout these 4 years.

I am particular grateful for the support of my mother, father and sister over the years. Only with their support I was able to attend University.

Finally, I would like to acknowledge the chance that was given to me by ITC, the European Commission (funding institution of the INACHUS project) and, Prof. dr. Norman Kerle and Prof. dr. Markus Gerke, for providing this PhD opportunity.

Table of Contents

List of figures

List of tables

1 Introduction

1.1 *Earthquakes: human, social and economic losses*

The United Nations Department of Economics and Social Affairs (UNDESA) in their 2014 report on World Urbanization Prospects, indicated that more than half (54%) of the world population is living in urban centers and that by 2050 this value will be situated around 66%. Already in 1999, Mitchell addressed the trend towards the increasing exposure to hazards, especially in megacities (cities with more than 10 million inhabitants) which are not prepared for such events (Mitchell, 1999). This increase in population exposure to hazards, among them earthquakes (see Figure 1), makes the disaster related field of growing importance. From the disaster risk to the disaster management all these fields have as objective to reduce the negative impact of such events.

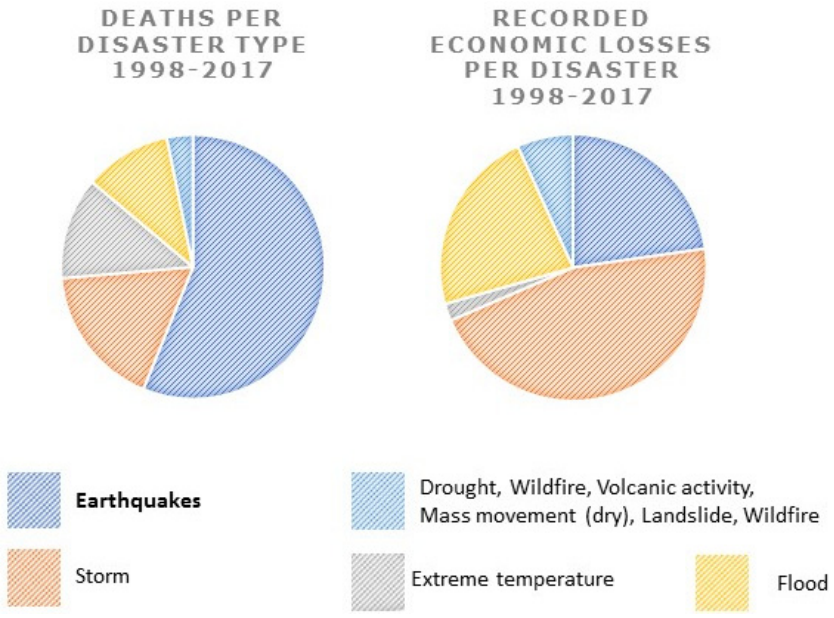

Figure 1 Relative death and recorded losses per disaster type– adapted from (Wallemacq and House, 2018)

Within disaster management, the disaster response phase is defined by the United Nations Office for the Disaster Risk Reduction (UNISDR) as "the provision of emergency services and public assistance during or immediately after a disaster in order to save lives, reduce health impacts, ensure public safety and meet the basic subsistence needs of the people affected". This definition clearly implies that the rescue operations by Urban Search and Rescue (USaR) and First Responders (FR) are one of the most important components of the disaster response phase since they are related with the task

of saving human lives. Performed by FR and USaR teams, these operations are time expensive since they are performed at a damaged building level and in a chaotic environment. Hence, prioritization of locations of where to deploy FR and USaR teams becomes a very important task.

This optimization effort is directly related with the detection of the most affected building blocks given that partial and totally collapsed buildings are a proxy for victim localization. The elapsed time between the event and the localization of collapsed buildings is of utmost importance in this phase; given the critical conditions of trapped victims.

Then a more detailed and qualitative assessment of damage is needed in the rehabilitation and recovery phase too. This phase focuses on the restoration of both services/facilities and living conditions of a given region and affected communities. For example, insurance companies need a detailed and accurate description of the damages of a given building; while local authorities need to assess the number of persons that need to be relocated to new housing. Such tasks need to move on to a more detailed damage assessment, where for example the façades are also considered.

Remote sensing images represent the conventional data source to determine the location and severity of damages over a region after a disastrous event (Dong and Shan, 2013). Most of the remote sensing platforms have been used for building damage assessment at several scales (Balz and Liao, 2010; Murtiyoso et al., 2014; Sui et al., 2014). The objectives vary according to the characteristics of both the sensor and platform used, and the desired application.

1.2 *Remote sensing imagery for the localization of partially and totally collapsed buildings*

There is a wide range of literature which focused on the mapping of partially and totally collapsed buildings, from satellite systems (Miura et al., 2007; Ural et al., 2011; Yusuf et al., 2001), traditional airborne systems (Fukuoka and Koshimura, 2012; Hasegawa et al., 2000; Sirmacek and Unsalan, 2009), unmanned aerial systems (Fernandez Galarreta et al., 2015)) or even terrestrial imaging systems (Armesto-González et al., 2010; Curtis and Fagan, 2013).

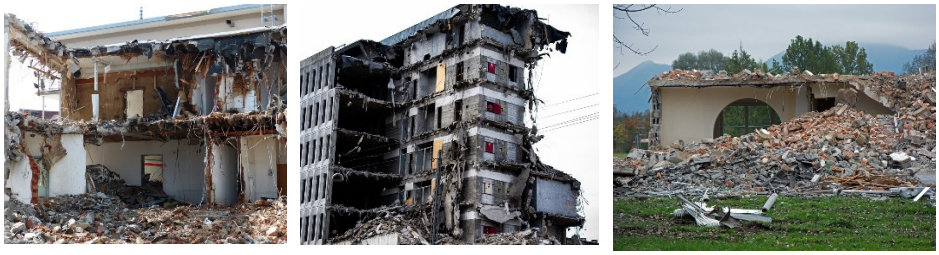

Figure 2 Examples of partially collapsed (2 left images) and total collapse, right

Satellite optical imagery is often used for synoptic damage assessment (Miura et al., 2007; Tong et al., 2012). The current high spatial resolution of satellite optical images (for example WorldView-3 with resolutions ~ 0.35m) may enable a per building damage assessment, while covering large areas. Copernicus, through its Emergency Management Service (EMS), and the Disaster Charter, are two agencies which currently use such satellite imagery to manually generate grading damage maps right after a given disaster. Hence, there has been an increasing amount of studies reporting on the automation of damage assessment from satellite optical images. Such approaches might rely on post-event data only (Dell'Acqua and Polli, 2011), pre- and post-event image data (Miura et al., 2007) and even considering height information retrieved from stereo pairs generated from the satellite images (Tong et al., 2012). Post-event only approaches usually rely on the radiometric features of the satellite images (Vetrivel et al. 2016) and/or alongside the height information (Tong et al., 2012). However, approaches considering pre-event images can further aid in the disambiguation between damaged and not damaged regions (Dong and Shan, 2013). However, the low resolution and nadir constrained view of satellite images may limit it to: 1) e.g. differentiate cluttered urban areas (such as narrow streets, slums, etc.) from damaged regions, 2) have a more detailed damage assessment regarding the damage state of a building (e.g. also considering façades).

Recent literature also used aerial manned platforms to survey regions where a disaster occurred (Corbane et al., 2011; Saito et al., 2010). This increased the resolution of the imagery collected to a decimetre level and at the same time allowed to capture oblique views, while having a lower coverage when compared with satellite. This increase in resolution and the ability to capture oblique views is advantageous twofold: while the increase in resolution allows to reduce the ambiguity between damaged and not damaged buildings (Booth et al., 2011; Kerle, 2010), the oblique views allow the façades to be assessed for damage (Booth et al., 2011; Mitomi et al., 2002; Saito et al., 2010). Given this, the EMS recently started signing contracts with private companies to acquire such aerial imagery after a disaster ("CGR supplies aerial survey to JRC for emergency," n.d.), as it happened already with the 2016 earthquakes in Italy. This interest in aerial imagery can also be noted on the several literature published regarding the use of such imagery for damage mapping in the last

couple of decades. Aerial television images captured with a tilt angle of about 30-45 degrees from the vertical direction were used to detect damages after the Kobe (Japan) earthquake in 1995 (Mitomi et al., 2002). The authors extracted several textures (e.g. co-occurrence matrix of edge intensity) from the video frames to determine the image characteristics of collapsed buildings. Using aerial systems consisting of five cameras (one nadir and one for each cardinal direction, Pictometry system), Saito et al. (2010) manually assessed the imagery to detect damaged buildings. The authors indicated that the visual interpretation of such images allowed to identify both collapsed and partially collapsed buildings and façade damages. Given the usual decimetre resolution of aerial surveys, object based image analysis started to be considered (Fukuoka and Koshimura, 2012; Li et al., 2011). In such cases, to consider groups of pixels was found more advantageous than pixel based approaches, given that with decimetre resolution objects in the scene (as well as damaged regions) were composed by a greater amount of pixels. To this regard texture features were found to be central for the damage identification. Following these works, Ma and Qin (2012) and Nex et al. (2014) also indicated that morphological features could complement the already rich information extracted using texture features. Gerke and Kerle (2011) extracted features from aerial oblique images and derived a 3D point cloud to detect damaged buildings after the 2010 Haiti earthquake. The authors considered three classes, based on the European macroseismic scale (Grünthal, 1998). Recently, 3D features and 2D CNN features were integrated by Vetrivel et al. (2017) using a multiple kernel learning, where the relevance of 2D CNN features was reported. This was mostly due to the often-noisy point clouds derived from dense image matching (Vetrivel et al. 2017) and the recent developments in computer vision and machine learning.

Unmanned aerial vehicles also started to be used to perform a more thorough damage assessment (Cusicanqui et al., 2018; Fernandez Galarreta et al., 2015). Such platforms have higher portability and higher resolution and incredible flexibility in terms of acquisitions when compared with the manned aerial platforms. Aerial manned platforms usually follow a predefined flight plan considering an oblique view for each cardinal direction plus the nadir captures. In this way, occlusions due to urban design are often present, especially considering old European city centres for example. Hence, the high portability of the UAV opens the possibility of directing the flights according to the needs of the user. These can focus the analysis on a specific set of buildings and be able to assess several building elements separately.

1.3 *Remote sensing imagery for the detection of façade damages*

The detection of partially and totally collapsed buildings from remote sensing images currently shows very promising results, mainly due to the higher accuracies achieved using state of the art image classification algorithms using CNN. However, constraining the damage detection to debris and rubble piles might leave out smaller damage evidences. Spalling, cracks and other smaller signs of damage are overlooked by approaches which were trained with image samples depicting debris and rubble piles, even if using oblique imagery (Vetrivel et al. 2017; Gerke and Kerle 2011), see Figure 3. Moving forward from the detection of rubble piles and debris and focusing on the façades gives more awareness to first responders regarding the damage state of a given region, where more damage information regarding the different elements of a building enables more informed decisions. Moreover, the detection of such smaller damage evidences is also useful for later stages of the disaster management cycle. Extended building damage catalogues are needed for the planning of recovery actions for example. Nonetheless, such extensive and comprehensive damage mapping relies on high-resolution and multi-view imagery given the often-smaller damage evidences. Airborne oblique imagery has been recently indicated to be promising to perform such assessments. Both manned and unmanned platforms have been used to assess the facades and perform more detailed damage assessments. Focusing on the specific façade damage detection Tu et al. (2017) took advantage of the symmetry often present in facades to determine damaged facades when that symmetry is not present, using only post-event image data collected from aerial manned platforms with decimetre level resolution. Fernandez Galarreta et al. (2015) used millimetre resolution imagery from UAV and terrestrial acquisitions to detect cracks on façades using the images and to detect slanted façades using the 3D point cloud.

Figure 3 Example of façade damages

1.4 *Research background, objectives and overall contributions*

The research reported on this thesis is part of the project *Technological and Methodological Solutions for Integrated Wide Area Situation Awareness and Survivor Localization to Support Search and Rescue Teams* (INACHUS), a 7th Framework Programme funded project (www.inachus.eu). The project aimed at time reduction related to the response phase of FR and USaR teams which translate in a higher number of victims rescued. A large consortium (20 partners from 10 EU countries) from several technological and scientific domains was needed to achieve such goal. An operations framework was established covering a broad set of stages, from the inference on the location of victim hotspots at a regional level, up to the localisation of the victims inside a damaged building. Three main fields were addressed: 1) simulation tools, incorporating wide-area hazard simulation and building collapse simulations; 2) remote sensing, making use of both passive and active sensors being airborne and terrestrial for the detection of building damages at several scales (comparing with the simulations in 1)); 3) human presence detection, such as a robot snake mounted with human detection sensors and mobile phones detection. These three main fields had to be integrated in a seamless manner also considering other parallel aspects, such as training material, ethics and standardization issues.

Overall, a wide-area damage assessment was coupled with dasymetric mapping in order to identify the regional hotspots. Earth observation tools such as UAVs equipped with both passive and active sensors were used to survey the disaster area and detect damaged buildings and their degree of destruction. Collapse simulation tools enabled both an early comprehension of the disaster magnitude and the understanding of the collapse itself. Nevertheless, this remote sensing and building collapse simulation tools could only infer, not detect, on the location of the entrapped victims both at a building block and building level. This was performed by other partners. The broadness of technological fields being tremendous accounted for the large consortium. ITC along with three more project partners addressed the remote sensing slice of the project. Specifically, ITC covered the use of multi-temporal and multi-resolution remote sensing imagery in the detection of building damages, namely partially and totally collapsed buildings, and façade damages. In accordance with the project this thesis focus on these two distinct subjects: detection of partially and totally collapsed structures, and façade damages. The latter is comprised by two parts, one focusing on the optimization of a façade damage detection procedure using the UAV and a second part aiming at a multi-temporal approach for the detection of façade damages using aerial manned platforms.

As indicated in the previous sub-section, the mapping of partially and totally collapsed buildings from remote sensing imagery is an extensively studied subject. To this regard several methods have been proposed. Such methods are usually related with the chosen platform to perform the aerial surveys, given the differences in resolution and view angle as well as image quality. The proposed frameworks are specifically designed for a given system (combination of a given platform with a given sensor, e.g. satellite optical imagery). This makes the approaches dependent on the amount of image data available for a given system, in order to be successful. This is more critical given the current state-of-the-art in image recognition tasks, where convolutional neural networks often need large amounts of image samples in order to achieve recognition capabilities. The developed algorithms have been conceived to cope on one hand with the lack of extensive datasets and, on the other hand, to use as input all the available images, regardless the used system (satellite or airborne). The main objective regarding the first part of the presented research is to assess how the combination of damaged image samples coming from satellite and aerial (manned and unmanned platforms) impact the image classification of debris and rubble piles of a given system. Several experiments are performed in order to assess the optimal fusion of such multi-resolution and multi-platform imagery. Specifically, these present the context, novelty and experiments regarding the use of multi-resolution optical image data for the detection of building damages, namely partially and totally collapsed structures.

The detection of rubble piles and debris is useful to identify partially or totally collapsed structures, however it leaves out several damage evidences. This is a different task to the mapping of rubble piles and debris since façade damages often entail several typologies of damage, from collapsed portions of the façade, to cracks on the walls. The few existing approaches either assume that façades often present symmetries or follow rule-based approaches specific for a given dataset. Hence, the second part of the research reported in this thesis focuses on the detection of damaged façades using aerial oblique imagery, being captured from unmanned (UAV) or manned vehicles. Within this broader subject, the approach using UAV focused on the efficiency of the damage detection approach, given its possible use by FR. Instead of running a damage detection algorithm on all the UAV high resolution images, the objective was to direct all these computations to the façade image patches. Hence, the objective was to extract only the façades from the wide range of images and then apply the damage detection algorithm to those specific image regions. A further study regarding façade damage detection was then performed. This focused on the mapping of façade damages using aerial manned platforms, given the higher areal coverage when compared with UAV and also due to the fact that these types of aerial oblique surveys are increasingly more common, especially in urban areas due to their ability to survey the façades. To this regard, the second part of the broader façade damage detection subject

focused on a multi-temporal approach, since such approach was still not considered for the specific case of façade damage detection. The recent interest in aerial oblique images (Vetrivel et al. 2017; Nyaruhuma et al. 2012; Murtiyoso et al. 2014), especially from grading damage map producers such as the EMS, we can expect pre-event imagery being available and used for the detection of façade damages alongside the post-event data. In this research several multi-temporal approaches are tested for the image classification of façade damages using aerial oblique imagery. The experiments focus on two different issues: 1) merging of pre- and post-event imagery within a CNN framework, 2) take advantage of the usual high overlap of aerial oblique image surveys to acquire several different views from the same façade and embed this in the proposed frameworks.

1.5 *Structure of the thesis*

This dissertation is composed of 7 chapters. While chapter 1 and chapter 7 are respectively the introduction and synthesis, the remaining chapters are scientific chapters holding specific research objectives, methods, results, discussion and conclusions. An overall content per chapter is indicated in the following paragraphs:

1. **Introduction**: motivates remote sensing image-based damage detection from a broader context, presents the background regarding the mapping of debris/rubble piles and façade damages, lays out the research objectives and the overall contributions
2. **Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach**: The first set of experiments regarding the use of multi-resolution imagery was tested for the specific case of satellite images. The focus of the experiments was on the optimal merge of the different sets of images (satellite and aerial, manned and unmanned). Different ways of merging this multi-resolution feature information were tested.
3. **Multi-resolution feature fusion for the image classification of building damages using convolutional neural networks**: This chapter is an extension of the previous one, where the multi-resolution approaches are applied to satellite and aerial (manned and unmanned) imagery. Furthermore, in this extended study, the geographical transferability was also tested for each of the different resolutions. Overall there is an improvement in the detection of damages when using multi-resolution approaches. This was more critical for the satellite and UAV case, while for the aerial manned case, a traditional approach was preferable.
4. **Towards a more efficient detection of earthquake induced façade damages using UAV oblique imagery**: This chapter is the first of three focusing on the façade damage detection. Specifically, this chapter focuses on providing a more efficient detection of façade damages when using UAV.

This is intended to be used by USaR in the field when needing to survey a building block. The objective was to direct all the damage computations to the images and image regions which contained façades, hence reducing the time needed for the running of damage algorithms on the whole set of multi-view images. However, it was noted that using a network trained on rubble piles and debris might not be optimal for the specific case of façade damage detection.

5. **Potential of multi-temporal oblique airborne imagery for structural damage assessment**: This chapter presents early results on the detection of façade damages using multi-temporal aerial oblique imagery. The general approach aimed at comparing the correlation coefficient between the pre- and post-event rectified façade image patches and two pre-event views of the same façade. While the correlation coefficient between epochs was much lower when the façade was damaged, it relied on the definition of a threshold to differentiate between intact and damaged façades.
6. **Image classification of façade damages using multi-temporal aerial oblique imagery**: This chapter is an extension of the previous one. Moving forward from rule based approaches, the focus of this chapter was on the optimal merge of pre- and post-event imagery within a deep learning approach. Moreover, given that in aerial oblique surveys a façade is observed from different views, this information was embedded in the framework merging both epochs feature information. Comparing with mono-temporal approaches there was a clear improvement. Furthermore, to consider several views per façade within a late fusion approach was preferable.
7. **Synthesis**: The final chapter presents an overview of the findings reported in the previous chapters. Also presents the conclusions regarding said findings and recommendations for future research.

The chapters of this thesis are based on peer-reviewed journal and conference papers. These follow a gradual set of experiments regarding both the mapping of debris and rubble piles, and façade damages. Given the shared goal of the objectives between contributions, there is often an overlap when presenting the background and related works. The chapters being standalone was found preferable due to the possible interest on a single chapter, where the reader does not need to consult any other part of the thesis for its full comprehension.

1.6 *References of the Introduction*

Armesto-González, J., Riveiro-Rodríguez, B., González-Aguilera, D., Rivas-Brea, M.T., 2010. Terrestrial laser scanning intensity data applied to damage detection for historical buildings. J. Archaeol. Sci. 37, 3037–3047. https://doi.org/10.1016/j.jas.2010.06.031

Balz, T., Liao, M., 2010. Building-damage detection using post-seismic high-resolution SAR satellite data. Int. J. Remote Sens. 31, 3369–3391. https://doi.org/10.1080/01431161003727671

Booth, E., Saito, K., Spence, R., Madabhushi, G., Eguchi, R.T., 2011. Validating Assessments of Seismic Damage Made from Remote Sensing. Earthq. Spectra 27, S157–S177. https://doi.org/10.1193/1.3632109

CGR supplies aerial survey to JRC for emergency [WWW Document], n.d. . CGR Spa. URL http://www.cgrspa.com/news/cgr-fornira-il-jrc-con-immagini-aeree-per-le-emergenze/ (accessed 11.9.15).

Corbane, C., Saito, K., Dell'Oro, L., Bjorgo, E., Gill, S.P.D., Emmanuel Piard, B., Huyck, C.K., Kemper, T., Lemoine, G., Spence, R.J.S., Shankar, R., Senegas, O., Ghesquiere, F., Lallemant, D., Evans, G.B., Gartley, R.A., Toro, J., Ghosh, S., Svekla, W.D., Adams, B.J., Eguchi, R.T., 2011. A comprehensive analysis of building damage in the 12 January 2010 Mw7 Haiti earthquake using high-resolution satellite and aerial imagery. Photogramm. Eng. Remote Sens. 77, 997–1009. https://doi.org/10.14358/PERS.77.10.0997

Curtis, A., Fagan, W.F., 2013. Capturing damage assessment with a spatial video: an example of a building and street-scale analysis of tornado-related mortality in Joplin, Missouri, 2011. Ann. Assoc. Am. Geogr. 103, 1522–1538. https://doi.org/10.1080/00045608.2013.784098

Cusicanqui, J., Kerle, N., Nex, F., 2018. Usability of aerial video footage for 3D-scene reconstruction and structural damage assessment. Nat. Hazards Earth Syst. Sci. Discuss. 1–23. https://doi.org/10.5194/nhess-2017-409

Dell'Acqua, F., Polli, D.A., 2011. Post-event only VHR radar satellite data for automated damage assessment. Photogramm. Eng. Remote Sens. 77, 1037–1043. https://doi.org/10.14358/PERS.77.10.1037

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS J. Photogramm. Remote Sens. 84, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011

Fernandez Galarreta, J., Kerle, N., Gerke, M., 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Nat. Hazards Earth Syst. Sci. 15, 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

Fukuoka, T., Koshimura, S., 2012. Quantitative analysis of tsunami debris by object-based image classification of the aerial photo and satellite image. J. Jpn. Soc. Civ. Eng. Ser B2 Coast. Eng. 68, I_371-I_375. https://doi.org/10.2208/kaigan.68.I_371

Gerke, M., Kerle, N., 2011. Automatic structural seismic damage assessment with airborne oblique Pictometry© imagery. Photogramm. Eng. Remote Sens. 77, 885–898. https://doi.org/10.14358/PERS.77.9.885

Grünthal, G., 1998. European Macroseismic Scale 1998 (EMS-98). , 99 pp., 1998. Centre Européen de Géodynamique et de Séismologie, Luxembourg.

Hasegawa, H., Aoki, H., Yamazaki, F., Matsuoka, M., Sekimoto, I., 2000. Automated detection of damaged buildings using aerial HDTV images. IEEE, pp. 310–312. https://doi.org/10.1109/IGARSS.2000.860502

Kerle, N., 2010. Satellite-based damage mapping following the 2006 Indonesia earthquake—How accurate was it? Int. J. Appl. Earth Obs. Geoinformation 12, 466–476. https://doi.org/10.1016/j.jag.2010.07.004

Li, X., Yang, W., Ao, T., Li, H., Chen, W., 2011. An improved approach of information extraction for earthquake-damaged buildings using high-resolution imagery. J. Earthq. Tsunami 05, 389–399. https://doi.org/10.1142/S1793431111001157

Ma, J., Qin, S., 2012. Automatic depicting algorithm of earthquake collapsed buildings with airborne high resolution image. IEEE, pp. 939–942. https://doi.org/10.1109/IGARSS.2012.6351400

Mitchell, J.K., 1999. Megacities and natural disasters: a comparative analysis*. GeoJournal 49, 137–142.

Mitomi, H., Matsuoka, M., Yamazaki, F., 2002. Application of automated damage detection of buildings due to earthquakes by panchromatic television images. Presented at the The 7th US National Conference on Earthquake Engineering.

Miura, H., Yamazaki, F., Matsuoka, M., 2007. Identification of damaged areas due to the 2006 central Java, Indonesia earthquake using satellite optical images. IEEE, pp. 1–5. https://doi.org/10.1109/URS.2007.371867

Murtiyoso, A., Remondino, F., Rupnik, E., Nex, F., Grussenmeyer, P., 2014. Oblique aerial photography tool for building inspection and damage assessment, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 309–313. https://doi.org/10.5194/isprsarchives-XL-1-309-2014

Nex, F., Rupnik, E., Toschi, I., Remondino, F., 2014. Automated processing of high resolution airborne images for earthquake damage assessment, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 315–321. https://doi.org/10.5194/isprsarchives-XL-1-315-2014

Nyaruhuma, A.P., Gerke, M., Vosselman, G., Mtalo, E.G., 2012. Verification of 2D building outlines using oblique airborne images. ISPRS J. Photogramm. Remote Sens. 71, 62–75. https://doi.org/10.1016/j.isprsjprs.2012.04.007

Saito, K., Spence, R., Booth, E., Madabhushi, G., Eguchi, R., Gill, S., 2010. Damage assessment of Port -au-Prince using Pictometry, in: 8th International Conference on Remote Sensing for Disaster Response. Tokyo Institute of Technology.

Sirmacek, B., Unsalan, C., 2009. Damaged building detection in aerial images using shadow Information. IEEE, pp. 249–252. https://doi.org/10.1109/RAST.2009.5158206

Sui, H., Tu, J., Song, Z., Chen, G., Li, Q., 2014. A novel 3D building damage detection method using multiple overlapping UAV images. ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XL–7, 173–179. https://doi.org/10.5194/isprsarchives-XL-7-173-2014

Tong, X., Hong, Z., Liu, S., Zhang, X., Xie, H., Li, Z., Yang, S., Wang, W., Bao, F., 2012. Building-damage detection using pre- and post-seismic high-resolution satellite stereo imagery: A case study of the May 2008 Wenchuan earthquake. ISPRS J. Photogramm. Remote Sens. 68, 13–27. https://doi.org/10.1016/j.isprsjprs.2011.12.004

Tu, J., Sui, H., Feng, W., Sun, K., Xu, C., Han, Q., 2017. Detecting building façade damage from oblique aerial images using local symmetry feature and the Gini Index. Remote Sens. Lett. 8, 676–685. https://doi.org/10.1080/2150704X.2017.1312027

Ural, S., Hussain, E., Kim, K., Fu, C.-S., Shan, J., 2011. Building extraction and rubble mapping for city Port-au-Prince post-2010 earthquake with GeoEye-1 imagery and lidar data. Photogramm. Eng. Remote Sens. 77, 1011–1023. https://doi.org/10.14358/PERS.77.10.1011

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2017. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Vetrivel, A., Kerle, N., Gerke, M., Nex, F., Vosselman, G., 2016. Towards automated satellite image segmentation and classification for assessing disaster damage using data-specific features with incremental learning, in: GEOBIA 2016. GEOBIA 2016, Enschede, The Netherlands. https://doi.org/10.3990/2.369

Wallemacq, P., House, R., 2018. Economic losses, poverty & disasters: 1998-2017.

Yusuf, Y., Matsuoka, M., Yamazaki, F., 2001. Damage assessment after 2001 Gujarat earthquake using Landsat-7 satellite images. J. Indian Soc. Remote Sens. 29, 17–22. https://doi.org/10.1007/BF02989909

2 Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach[1]

[1] This chapter is based on the article:
Duarte, D., Nex, F., Kerle, N., and Vosselman, G.: Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-2, 89-96, https://doi.org/10.5194/isprs-annals-IV-2-89-2018, 2018.

Abstract

The localization and detailed assessment of damaged buildings after a disastrous event is of utmost importance to guide response operations, recovery tasks or for insurance purposes. Several remote sensing platforms and sensors are currently used for the manual detection of building damages. However, there is an overall interest in the use of automated methods to perform this task, regardless of the used platform. Owing to its synoptic coverage and predictable availability, satellite imagery is currently used as input for the identification of building damages by the International Charter, as well as the Copernicus Emergency Management Service for the production of damage grading and reference maps. Recently proposed methods to perform image classification of building damages rely on convolutional neural networks (CNN). These are usually trained with only satellite image samples in a binary classification problem, however the number of samples derived from these images is often limited, affecting the quality of the classification results. The use of up/down-sampling image samples during the training of a CNN, has demonstrated to improve several image recognition tasks in remote sensing. However, it is currently unclear if this multi resolution information can also be captured from images with different spatial resolutions like satellite and airborne imagery (from both manned and unmanned platforms). In this chapter, a CNN framework using residual connections and dilated convolutions is used considering both manned and unmanned aerial image samples to perform the satellite image classification of building damages. Three network configurations, trained with multi-resolution image samples are compared against two benchmark networks where only satellite image samples are used. Combining feature maps generated from airborne and satellite image samples, and refining these using only the satellite image samples, improved nearly 4% the overall satellite image classification of building damages.

2.1 Introduction and related work

Building damage maps have been recurrently used in the response and recovery phase of the disaster management cycle. Damaged buildings may be a proxy for victim localization (Dell'Acqua and Gamba, 2012) and their identification can also aid to plan and delineate recovery activities (Eguchi et al., 2009). Remote sensing has been extensively used to perform the damage assessment of a given region affected by a disastrous event (Dell'Acqua and Gamba, 2012; Dong and Shan, 2013; Gerke and Kerle, 2011; Vetrivel et al., 2017). The platforms used in remote sensing usually have a wide coverage, fast deployment and high temporal frequency while the collected data allow to automate building damage assessment procedures (Ural et al., 2011).

A wide variety of remote sensing sensors mounted on different platforms have been used to map building damages (Armesto-González et al., 2010;

Dell'Acqua and Polli, 2011; Gokon et al., 2015; Khoshelham et al., 2013; Marin et al., 2015; Vetrivel et al., 2017). However, there has been a growing interest regarding the use of images (Curtis and Fagan, 2013; Fernandez Galarreta et al., 2015; Vetrivel et al., 2015, 2016a, 2017).

In this regard, synoptic satellite imagery can be readily available and provide the first overview over a region struck by a disastrous event such as an earthquake (Dell'Acqua and Gamba, 2012). The International Charter (IC) (Bessis et al., 2004) and the Emergency Management Service (EMS) (Copernicus programme, European Commission), are two institutions which use such imagery to provide geoinformation to regions affected by disasters. The IC and EMS mostly rely on the manual interpretation of satellite images to identify damaged buildings, despite the amount of proposed automated methods. However, scene characteristics, cloud cover, limited resolution and viewpoint, limited time by map producers to develop new operational methods; hinder the automation of these procedures (Kerle, 2010; Vetrivel et al., 2016a).

Other platforms coupled with cameras have also been used to map damages (Sui et al., 2014; Vetrivel et al., 2016b). Manned and unmanned aerial vehicles (UAV) enable the acquisition of images at a higher-resolution and can also perform oblique flights, introducing another level of damage information regarding the façades (Tu et al., 2017). In this regard, the Joint Research Center (JRC, European Commission) awarded a contract in 2015 to a consortium of private companies for the provision of aerial imagery after a disastrous event within a European context ("CGR supplies aerial survey to JRC for emergency," n.d.). UAV images have become a normal source of information for many rescue teams in the recent earthquakes in Nepal (2015) and Italy (2016). These trends have pushed many researchers (Duarte et al., 2017; Sui et al., 2014; Vetrivel et al., 2017) to develop damage detection algorithms exploiting these high-resolution images.

The use of overlapping images may allow the generation of 3D point clouds through dense image matching. The set of geometrical information extracted from point clouds can be used alongside the images for the detection of building damages (Fernandez Galarreta et al., 2015; Vetrivel et al., 2017). Their added value can be marginal if single epoch data are considered (Duarte et al., 2017; Vetrivel et al., 2017). Furthermore, the generation of 3D point clouds is still very time consuming, hindering their use in early response tasks. The quality of these 3D data is directly related with the resolution of the input images, which limits the use of the 3D generated from satellite imagery.

The achieved results regarding the use of airborne and UAV images are promising and their use is drastically increasing in recent years. However, satellite images are still the first and most common source for damage

assessment. For this reason, a more reliable method to automate the detection from these images would be needed.

The most recent approaches to perform satellite image classification of building damages use CNN (Vetrivel et al., 2016a). The used networks are very similar to the ones used in the computer vision domain (Krizhevsky et al. 2017). Satellite image samples are used for the training of the network, in a binary classification scheme (i.e. damaged and not damaged areas). However, the number of samples from satellite images is relatively small, while a wide variety of images acquired with airborne platforms, both manned and unmanned, are available too. These data are currently used to train a network which classifies images with the same resolution (Vetrivel et al., 2017). In computer vision and remote sensing, the use of multi-resolution data has demonstrated to improve the overall image classification and segmentation (Fu et al., 2017; Hamaguchi et al., 2017; Lin et al., 2016; Liu et al., 2016). The multi-resolution training is usually performed artificially (Fu et al., 2017; Hu et al., 2015; Li et al., 2015; Shen et al., 2015; Tang and Mohamed, 2012), up/down sampling the images at several scales. However, a multi-resolution approach using image data from different platforms and sensors has not been tested yet.

The aim of this chapter is to assess if the combined use of different resolution images improves the image classification of building damages from satellite images using CNN (Figure 1).

The main idea is that the native multi-resolution information of remote sensing imagery (i.e. satellite and airborne) can be captured by a CNN, improving the satellite image classification. Several CNNs configurations have been tested to assess how the image samples from different resolutions can influence the performance of the classification of building damages. Two recent developments in the computer vision domain are used: residual connections and dilated convolutions. More details regarding the developed approach are described in Section 2. This is then followed by an experiments section (3) which details the datasets (Section 3.1) used to test the approach, presents the experiments (Section 3.2) and the achieved results (Section 3.3). The discussion and the conclusions are finally given in Section 4 and Section 5 respectively.

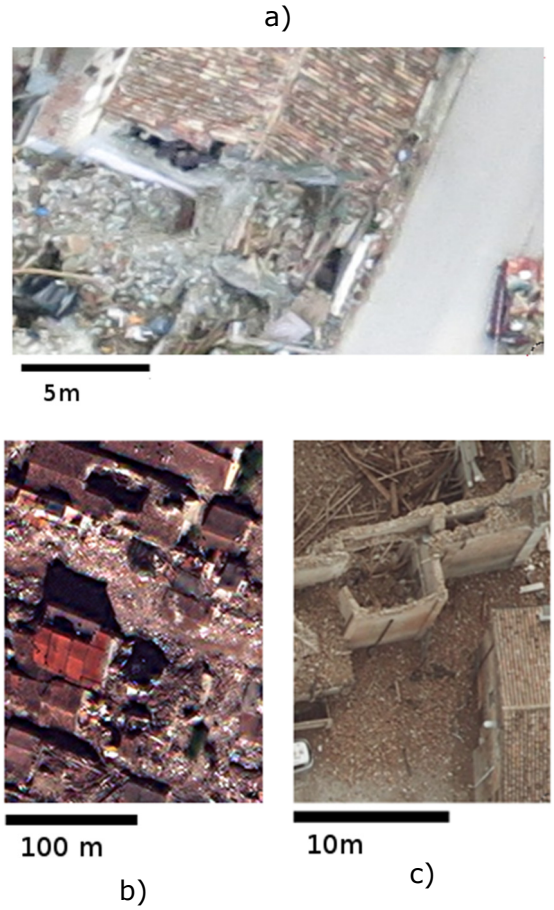

Figure 4 Examples of damaged and undamaged regions in a) UAV (Pescara del Tronto,Italy, 2016), b) satellite (WorldView 3, Amatrice, Italy, 2016) and c) manned aerial vehicles (St Felice, Italy, 2012) imagery.

The main idea is that the native multi-resolution information of remote sensing imagery (i.e. satellite and airborne) can be captured by a CNN, improving the satellite image classification. Several CNNs configurations have been tested to assess how the image samples from different resolutions can influence the performance of the classification of building damages. Two recent developments in the computer vision domain are used: residual connections and dilated convolutions. More details regarding the developed approach are described in Section 2. This is then followed by an experiments section (3) which details the datasets (Section 3.1) used to test the approach, presents the experiments (Section 3.2) and the achieved results (Section 3.3). The discussion and the conclusions are finally given in Section 4 and Section 5 respectively.

2.2 *Methodology*

Five different CNN architectures are defined. Two are used as benchmark and the remaining three are used to test the multi-resolution approach. Regarding the benchmark networks, the first is trained from scratch and the other one is fined-tuned on the generic satellite image samples provided by Cheng et al. (2017). The three multi-resolution test networks have been conceived to analyze the best way to combine and exploit features from each image resolution level.

All the networks take advantage of residual connections and dilated convolutions. This section explains these two central components of the networks while the two basic modules of the networks are then described in Section 2.1. The networks architectures used in the tests are finally presented in Section 3.

Residual connections: The depth of CNN have shown an increase in their capabilities to retrieve relevant information from images (Telgarsky, 2016). The usual hierarchical stacking of convolutional layers allows the network to learn from lower level features to higher levels of abstraction. Nonetheless, a given layer *l* may need feature information not only from the layer *l-1* but also from other previous layers (*l-2*, etc.). Residual connections (He et al., 2016) enable this process, by feeding a given layer to the previous one, as in the classical hierarchical approach, summed with a given output of earlier layers (Figure 2). In this way, every level of a given residual network effectivilly contributes to the final recognition task. Figure 2 shows a scheme of a residual connection and its interactions within a network. In this approach, features are extracted from remote sensing imagery at different spatial resolutions, where the relevance and complexity of a given feature may vary between the considered resolution levels. Thus, it is mandatory to capture and retain all of these levels of feature complexity through the use of residual connections.

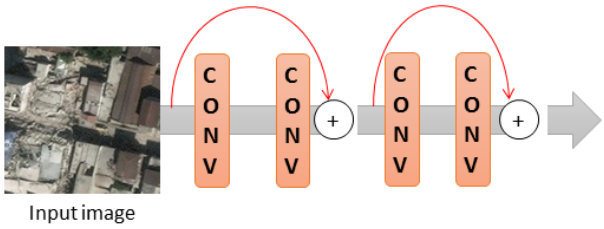

Figure 5 Simple scheme of possible residual connections within a CNN. The grey arrow shows a classical approach, while the red arrows show the new added (residual) connections.

Dilated convolutions: Another central aspect of a network capable of capturing multi-resolution information is its ability to capture spatial context. Recently, Yu and Koltun (2016) proposed the use of dilated convolutions (Figure 3) in CNN. These dilated convolutions consist of convolutions applied

to a given input image with a kernel having defined gaps (Figure 3). The receptive field of the network is bigger, capturing more contextual information (Hamaguchi et al., 2017). These dilated convolutions allow the integration of knowledge of the wider context (Hamaguchi et al., 2017) and at the same time depict finer details (Yu and Koltun, 2016). This is especially relevant for a multi-resolution approach since several sizes of patterns at different resolutions may contribute to the classification task.

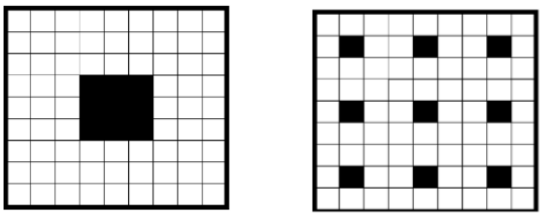

Figure 6 a) 3x3 kernel with dilation 1, b) 3x3 kernel with dilation 3

2.1.1 Basic convolutional set and modules definition:

The architecture of the CNN is composed by two main modules: 1) context module, followed by 2) resolution specific module (Figure 5). This structure was inspired by the works of Hamaguchi et al. (2017), Yu et al. (2017) and He et al. (2016). The general idea is that both context and resolution specific information is needed (Hamaguchi et al., 2017), hence the use of the two distinct modules.

Both modules are built stacking basic convolutional sets. These are composed of a convolution, batch normalization and ReLU (CBR, see Figure 4 a)) (He et al., 2016; Ioffe and Szegedy, 2015; Yu et al., 2017). Two basic convolutional sets bridged by a residual connection form a main CBR block, as shown in Figure 4 b). In each CBR, different number of filters and dilation values can be adopted. Both the context and resolution specific modules are composed of a sequence of CBRs with different numbers of filters and dilation rates as indicated in Figure 5.

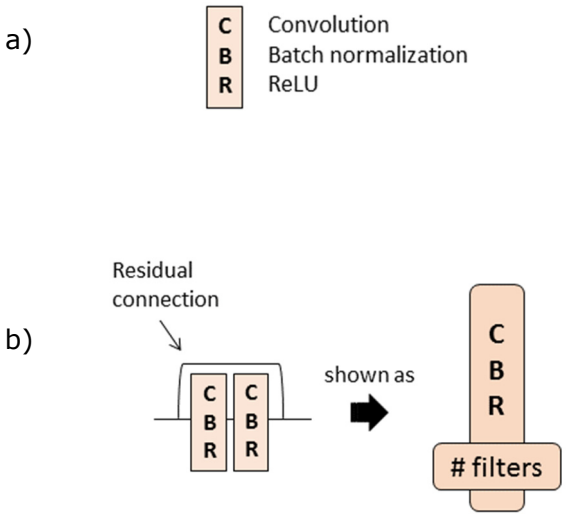

Figure 7 Basic convolutional set (a). Basic group of convolutions used to build the context and (b) resolution specific modules indicating the number of filters used

The context module (Figure 5 a)) is composed of several stacked CBRs with increasing dilation and increasing number of filters, with the objective of gradually capturing larger feature representations (Hamaguchi et al., 2017; Yu et al., 2017). The increasing number of filters over a CNN follows the state of the art approaches (He et al., 2016; Simonyan and Zisserman, 2015), more filters for higher level feature representation. The initial feature map is reduced from 224x224 (input) to 28x28px using a stride of 2, instead of 1 in the first three sets of CBRs. The use of larger stride has shown better performances than the max pooling operations, mainly because of the use of dilated convolutions (Yu et al., 2017). The kernel size of all the convolutions is 3x3 (Springenberg et al., 2015).

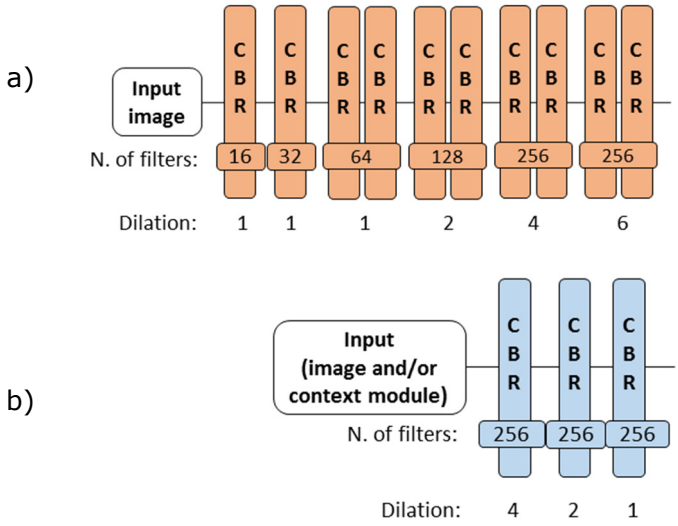

Figure 8 a) Context module, b) resolution specific module. Resolution specific module does not contain residual connections.

The increase in the dilation factor can create artificacts on the resulting feature maps, due to the gaps generated by the dilated kernel (Hamaguchi et al., 2017; Yu et al., 2017). To attenuate this drawback, the dilation increase in the context module is compensated in the resolution specific module with a gradual reduction of the dilation value and the removal of residual connections from the basic CBR blocks (Yu et al., 2017). This also allows to re-capture the more local features (Hamaguchi et al. 2017), which might be lost due to the increasing dilations in the context module.

For the classification part of the network, global average pooling followed by a convolution which maps the feature map size to the number of classes, is applied. Since this is a binary classification problem, a sigmoid function is used as activation.

2.3 *Experiments*

2.3.1 Dataset and training samples

There are two subsets of data: a) a multi-resolution dataset formed by three sets of images corresponding to satellite and airborne images (manned and UAV platforms) and b) a set of generic satellite image samples, which is used in one of the benchmark approaches.

Regarding the multi-resolution data, three sets of images, one set for each level of resolution, are considered: satellite, manned and unmanned aerial vehicles (Table 1). Most of the datasets depict real earthquake-induced building damages; however, there are also images from controlled demolitions.

The satellite images cover five different geographical locations in Italy, Ecuador and Haiti (Table 1). The satellite imagery was collected with WorldView 3 (Amatrice, Pescara del Tronto and Portoviejo) and GeoEye 1 (L'Aquila, Port-au-Prince). These data are pansharpened and have a variable resolution between 0.4 and 0.6m.

The airborne imagery consists of nadir and oblique imagery with a ground sampling distance (GSD) of 12-18 cm for the manned vehicles and of 2-10 cm for the UAV. The differences in image content at a given level of resolution (different illumination settings, view angles, sensors characteristics, morphology of buildings and urban landscape) are further increased by the multi-resolution aspect.

The samples are extracted for each resolution from the set of images indicated before. First, damaged and undamaged image regions are manually delineated, see Figure 6. Every cell that contains more than 60% of its area covered by one of the classes is cropped and used as an image sample for that same class. The grid size varies according to the resolution: satellite 80x80px, airborne (manned vehicles) 100x100px and airborne (UAV) 160x160px. The variable size of the image samples is set in order to keep in count the different resolution and the extension of the area captured in each patch. Due to the scarcity of satellite image samples (Table 1), to consider a smaller patch in this level of resolution, allowed to extract a higher number of samples.

Table 1 Overview of the location and quantity of satellite and airborne samples. The ++ locations indicate controlled demolitions of buildings.

Location	N. of samples		Month/Year of event
	Damaged	Not damaged	
	Satellite samples		
Aquila	115	118	April 2009
Port-au-Prince	732	701	January 2010
Portoviejo	147	163	April 2016
Amatrice	165	180	August 2016
Pesc. Tronto	93	94	August 2016
Total	**1252**	**1256**	
	Airborne (manned vehicles) samples		
Aquila,	336	385	April 2009
St Felice	587	593	May 2012
Amatrice	320	362	August 2016
Tempera	259	260	April 2009
Bidonville	229	229	January 2010
Port-au-Prince	749	712	January 2010
Onna	387	365	April 2009
Total	**2867**	**2906**	
	Airborne (UAV) samples		
Aquila	113	131	April 2009
Wesel	90	94	++
Portoviejo	216	208	April 2016
Pesc. Tronto	218	264	August 2016
Katmandu	309	288	April 2015
Taiwan	187	611	February 2016
Gronau	457	501	++
Mirabello	502	453	May 2012
Lyon	312	310	++
Total	**2704**	**2860**	

The number of samples is approximately the same for the damaged and undamaged classes. However, the number of samples is not balanced among the 3 levels of resolution. The number of satellite image samples is two-fold lower when compared to the other two levels of resolution.

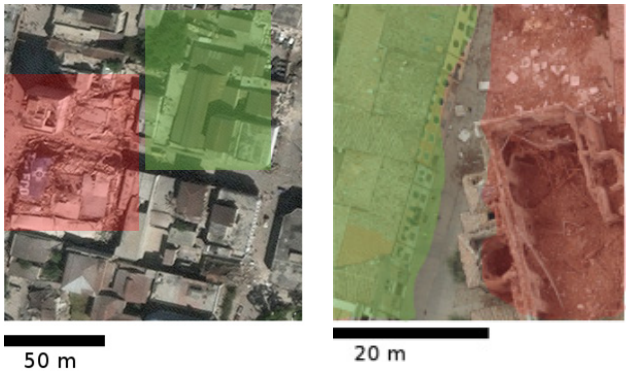

Figure 9 Examples of damaged (red) and non-damaged (green) areas digitized in satellite (GeoEye 1, Port-au-Prince, Haiti, 2010), left. Airborne (manned platform) (St Felice, Italy, 2012) imagery, right.

The generic satellite images samples are taken from a freely available benchmark dataset: NWPU-RESISC45 (Cheng et al., 2017). This benchmark dataset contains 45 classes with 700 satellite image samples per class. From these, fourteen classes were selected and divided into two broader classes, built and non-built (Table 2). Instead of considering the total 31500 samples, only fourteen classes are considered (9800) to reduce the computational cost of the approach.

Table 2 Fourteen classes of the benchmark dataset (NWPU-RESISC45) divided in built and non-built classes. Each class contains 700 samples, totaling 9800 image samples.

Built	Non-built
Airport	Beach
Commercial area	Circular farmland
Dense residential	Desert
Freeway	Forest
Industrial area	Mountain
Medium residential	Rectangular farm
Sparse residential	Terrace

2.3.2 Experiments

Using the modules defined before in section 2.2, five different networks are derived from the architectures shown in Figure 7. The first two networks are used as benchmarks for the other tests involving the multi-resolution architecture. In the first benchmark network (Figure 7 a)), the satellite training samples are fed into a network composed of the context module and the resolution specific module. The second benchmark uses the same architecture as defined in Figure 7 c) (mresB). It feeds the generic satellite image samples (Table 2) into the context module, while the resolution specific is only fed with the satellite samples. Due to the low number of damage domain satellite image samples (2508) when compared to the other levels of resolution (around 5700), training a network from scratch may not be optimal (Tajbakhsh et al.,

2016). For this reason, the second benchmark (henceforth referred as benchmark_ft), fine tunes the learned features from generic satellite samples, with damage domain specific satellite image samples.

The other three networks combine both the context and the resolution specific modules. The overall aim of these tests is to understand if sharing features between resolutions (Figure 7 b) and c)) captures more relevant information than merging the output of each separate context module (Figure 7 d)). A more detailed explanation of these three networks is given below:

mresA: feeds the training data of all resolutions to the context followed by the resolution specific module. In this way the extracted features of both modules are shared between resolutions, Figure 7 b).

mresB: all the training data of all resolutions are fed into the context module. However, the resolution specific module is only fed with the satellite samples. In this case the context module serves as base model with its weights that are tuned in the resolution specific module, Figure 7 c).

mresC: each data resolution is given to a different context module. The output of these modules is subsequently summed. These summed feature maps are used to initialize the resolution specific module that considers only satellite image samples, Figure 7 d).

The stockastic gradient descent (Wilson et al., 2017), with momentum of 0.9 and with a decreasing learning rate, is used in the optimization. The initial learning rate is of 10^{-2}, decreasing by a factor of 10 every 30 epochs (total of 120), with a weight decay of 10^{-2}. This is set for the benchmark and mresA networks. For the other two networks, the context and resolution specific modules are executed separately. In these cases, the context module is performed with the same learning rate parameters of the benchmark and mresA. However, the resolution specific learning rates differ. The mresB (and benchmark_ft) resolution specific module has the learning rate initially set at 10^{-3}, decreasing by a factor of 10 every 30 epochs, with a weight a decay of 10^{-6}. In the case of the mresC the learning rate is set initially to 10^{-4}, with the same decreasing rate and weight decay as mresB. These parameters are obtained empirically.

In the benchmark and mresA the networks are learning from scratch, hence the aggressive learning rate. While in the benchmark_ft,mresB and mresC, the resolution specific module intends to take advantage of the weights obtained by the context module, hence the lower learning rate parameters. In this way, the multi-resolution context information is refined for the specific case of the satellite image classification of building damages.

During the training of every network, data augmentation is performed since this has shown to avoid overfitting and improve the overall image classification (Krizhevsky et al., 2017; Simonyan and Zisserman, 2015). The used data

augmentation consists of random translations, rotations, image normalization and up/downsampling of the images. The networks were run for 120 epochs with a batch size of 8. The input size for the network is 224x224px. The image samples are zero padded to fit in this template, instead of being resized (Vetrivel et al., 2016a).

The training is performed using 70% of the samples of each resolution, while the validation uses 30% of the satellite image samples. This ratio is applied to each location separatly. The selected samples for both the training and validation remains the same for all the experiments.

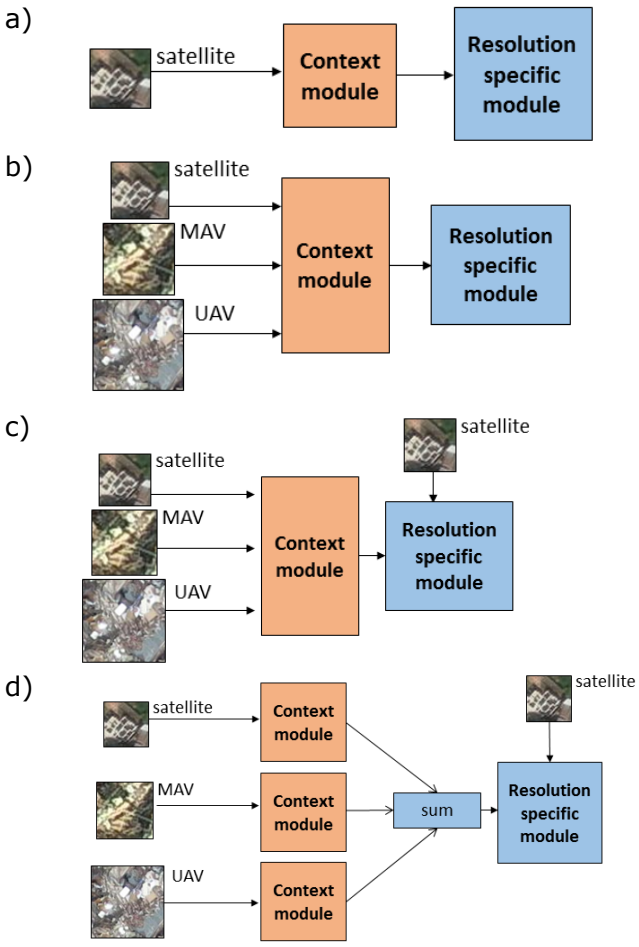

Figure 10 Tested network configurations: a) benchmark, b) multi-resolution A (mresA), c) multi-resolution B (mresB) and d) multi-resolution C (mresC). Details on the text.

2.3.3 Results

The achieved results of the use of the five network architectures are presented below in Table 2.

Table 3 Results of experiments

Network	Accuracy	Parameters	Training samples
benchmark	0.905	8.6M	1718
benchmark_ft	0.904	8.6M	11518
mresA	0.898	8.6M	8685
mresB	0.924	8.6M	8685
mresC	0.944	18.4M	8685

As indicated in this table, the benchmark network trained from scratch (benchmark) marginally outperforms the one which used generic satellite image samples in the context module and posteriorly fine-tuned it with the damage domain samples (benchmark).

Most of the multi-resolution approaches overcome the benchmark networks. Only mresA underperformed the two benchmark networks. The best performing network was mresC with an accuracy increase of almost 4% compared to the benchmark. This network also outperformed mresB by 2%. The network mresC is also the one with the higher number of parameters since 3 context modules were added before the resolution specific module. The number of training samples is of 1718 for the benchmark 11518 for the benchmark_ft and 8685 for the rest of the networks.

To better understand and validate the networks behaviour, a second test was conducted by feeding them with new and unused satellite image patches. These input patches were of 224x224 px (i.e. different from the sample sizes of 80x80 px). Figure 8 and Figure 9 show activations given by the last set of filters of all the multi-resolution networks and the benchmark one with the higher accuracy (benchmark, Table 2). In particular, for each network and from the set of 256 feature maps, the one with the higher average activation value is visualized.

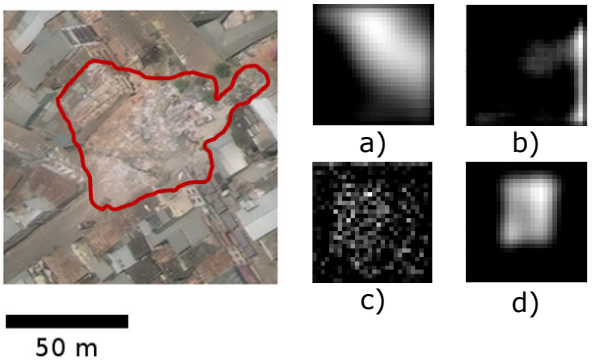

Figure 11 Satellite image sample (collected with WorldView-3, Porto Viejo, Ecuador, 2016), with damaged area manually outlined in red, fed into the network. Higher activation value of the last set of feature maps of the benchmark b), mresA c), mresB d) and mresC

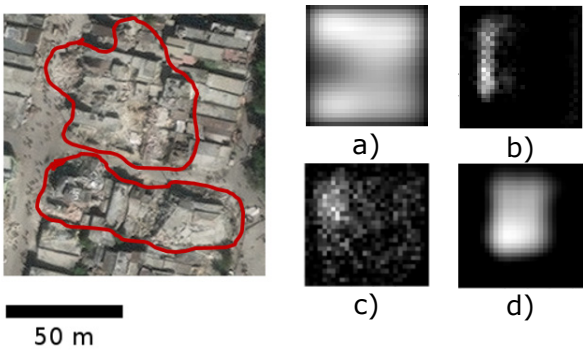

Figure 12 Satellite image sample, with the damage manually outlined in red (GeoEye 1, Port-au-Prince, Haiti, 2010) fed into the network. Higher activation value of the last set of feature maps of the benchmark a), mresA b), mresB c) and mresC d) networks

The activation from mresC (Figure 8 d)) shows a stronger agreement with the damaged area in red, when considering all the presented activations. However, smaller damaged areas are not considered as damaged. The activation from the benchmark (Figure 8 a)) also shows localization capabilities, but it is less discriminative in correspondence of non-damaged areas. Figure 8 b) presents the activation from mresA, where some difficulty to localize the damaged area from the given patch is evident. The mresB (Figure 8 c)), fails to localize the damage.

Another example is presented in Figure 9, left. In this case the mresC activation (Figure 9, d)), from the four activations, is the one that shows the better agreement with the damaged region. As in the previous case, there are smaller damage regions that are not identified in the activation. The benchmark (Figure 9, a)) activation goes across the whole image sample, including areas which are not damaged. mresA (Figure 9, b)) and mresB (Figure 9, c) only focus on the damaged area on the left upper part of the sample.

Both figures, mresA and mresB, present noisier activations than the benchmark and the mresC.

2.4 *Discussion*

The presented results indicate an improvement in the satellite image classification of building damages thanks to the use of different training samples from different spatial resolutions.

Only one multi-resolution network did not improve the classification accuracy compared to the used benchmarks. Two factors could have contributed to this: 1) this network was the only multi-resolution network where the resolution specific module was not trained only considering the satellite image samples; 2) the number of satellite training samples is twofold lower if compared with the other two resolutions. This might have led the networks to discard features which might be relevant for the satellite resolution.

The other two networks, which take input samples from all the resolution levels in their context module, outperform the benchmark tests. In this regard, the sum of the feature maps coming from the context module of each of the resolutions (mresC) seems to be more beneficial than feeding all of them into the same context module (mresB). In the case the context module is shared, the network might discard satellite features, due to an unbalanced number of training samples between the different image resolutions. This is in agreement with other remote sensing studies where the up/down sampled image samples are fed into a different network (or parts of the network) and each feature map is then summed to provide a stronger classifier (Fu et al., 2017; Maggiori et al., 2017). The number of parameters is also higher in the best performing resolution; this might have a positive effect on the performance.

Considering previous works (Vetrivel et al., 2016a), there was an increase (around 15%) in the accuracy of satellite image classification of building damages, even without considering the multi-resolution aspect. This accuracy difference is, however, closely related with recent advancements in the image classification algorithms using CNN (He et al., 2016; Krizhevsky et al., 2017).

The activation maps confirm the results provided by the accuracy assessment; also in this case mresC outperform the other methods. However, the activations of this network appear to be smoother; smaller signs of damage might not be considered. In contrast, the activation maps of networks which shared the context module present a noisier activation and seem to generate artefacts as indicated in Hamaguchi et al. (2017) and Yu et al. (2017), even after decreasing the dilation value in the resolution specific module.

The learning rate was found to be critical. The used parameters were tuned empirically and a small change in the parameter values showed to have a high

impact on the final result. The presented results represent the best accuracy values achieved with each network configuration.

2.5 *Conclusions and future developments*

This chapter assessed the combined use of remote sensing imagery with different resolutions within a CNN approach, to perform the satellite image classification of building damages.

The combined use of several resolutions and their different combination in the training of the CNN, improved the accuracy of satellite image classification of building damages by nearly 4%. The addition of feature maps from the different resolutions has shown to capture more relevant information than having these shared in a single network. The activations of the best performing network, which sums the feature maps coming from the several resolutions, have shown a better agreement with manually defined damaged regions. However, the activations also show that this network is not able to identify smaller signs of damage, which can be critical for any decision maker considering a damaged map generated by such an automated approach.

Since the shown results are only related with the overall accuracy and behaviour of the networks, more research is needed to assess in which specific conditions this multi-resolution approach improves damage mapping. The datasets used in this experiment mostly refer to the same geographical regions (Haiti and Italy) and the same disastrous events, which could be one of the reasons for the reported results.

With the expected increase in the amount of collected imagery from several different platforms (both manned and unmanned platforms), this multi-resolution aspect of CNN can be beneficial in many practical cases. The trained networks would be very useful in the damage assessment at regional level, where satellite images are currently the only used source of information. This model could be further refined adding location specific samples in an online learning approach (Vetrivel et al., 2016a). In an early post-disaster setting, this multi-resolution capability is even more meaningful, due to the different sources of imagery that might be collected. While satellite may be the first set of available data, there is a continuous capture of airborne multi-resolution data from the initial stages of the response phase.

New tests will be performed using the same number of samples for every resolution. This would allow to better understand the impact of using unbalanced number of data with different resolutions. The use of only airborne samples as training to classify damages from satellite imagery will be then considered in order to assess the transferability of learned features to different resolutions.

The successful use of multi-resolution remote sensing image samples should also be extended to other image classification problems with more classes. There is an increasing amount of multi-resolution image data available and, in that sense, a multi-resolution approach taking advantage of such large amount of data would be beneficial.

2.6 *References of Chapter 2*

Armesto-González, J., Riveiro-Rodríguez, B., González-Aguilera, D., Rivas-Brea, M.T., 2010. Terrestrial laser scanning intensity data applied to damage detection for historical buildings. Journal of Archaeological Science 37, 3037–3047. https://doi.org/10.1016/j.jas.2010.06.031

Bessis, J.-L., Béquignon, J., Mahmood, A., 2004. The International Charter "Space and Major Disasters" initiative. Acta Astronautica 54, 183–190. https://doi.org/10.1016/S0094-5765(02)00297-7

CGR supplies aerial survey to JRC for emergency [WWW Document], n.d. . CGR spa. URL http://www.cgrspa.com/news/cgr-fornira-il-jrc-con-immagini-aeree-per-le-emergenze/ (accessed 11.9.15).

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE 1–19. https://doi.org/10.1109/JPROC.2017.2675998

Curtis, A., Fagan, W.F., 2013. Capturing damage assessment with a spatial video: an example of a building and street-scale analysis of tornado-related mortality in Joplin, Missouri, 2011. Annals of the Association of American Geographers 103, 1522–1538. https://doi.org/10.1080/00045608.2013.784098

Dell'Acqua, F., Gamba, P., 2012. Remote sensing and earthquake damage assessment: experiences, limits, and perspectives. Proceedings of the IEEE 100, 2876–2890. https://doi.org/10.1109/JPROC.2012.2196404

Dell'Acqua, F., Polli, D.A., 2011. Post-event only VHR radar satellite data for automated damage assessment. Photogrammetric Engineering & Remote Sensing 77, 1037–1043. https://doi.org/10.14358/PERS.77.10.1037

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS Journal of Photogrammetry and Remote Sensing 84, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2017. Towards a more efficient detection of earthquake induced facade damages using oblique UAV imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W6, 93–100. https://doi.org/10.5194/isprs-archives-XLII-2-W6-93-2017

Eguchi, R.T., Huyck, C.K., Ghosh, S., Adams, B.J., McMillan, A., 2009. Utilizing new technologies in managing hazards and disasters, in: Showalter, P.S., Lu, Y. (Eds.), Geospatial Techniques in Urban Hazard and Disaster

Analysis. Springer Netherlands, Dordrecht, pp. 295–323. https://doi.org/10.1007/978-90-481-2238-7_15

Fernandez Galarreta, J., Kerle, N., Gerke, M., 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Natural Hazards and Earth System Science 15, 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sensing 9, 498. https://doi.org/10.3390/rs9050498

Gerke, M., Kerle, N., 2011. Automatic structural seismic damage assessment with airborne oblique Pictometry© imagery. Photogrammetric Engineering & Remote Sensing 77, 885–898. https://doi.org/10.14358/PERS.77.9.885

Gokon, H., Post, J., Stein, E., Martinis, S., Twele, A., Muck, M., Geiss, C., Koshimura, S., Matsuoka, M., 2015. A method for detecting buildings destroyed by the 2011 Tohoku earthquake and tsunami using multitemporal TerraSAR-X data. IEEE Geoscience and Remote Sensing Letters 12, 1277–1281. https://doi.org/10.1109/LGRS.2015.2392792

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., Hikosaka, S., 2017. Effective use of dilated convolutions for segmenting small object instances in remote sensing images arXiv:1709.00179.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. IEEE, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing 7, 14680–14707. https://doi.org/10.3390/rs71114680

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Presented at the 34th International Conference on Machine Learning, Sydney, Australia.

Kerle, N., 2010. Satellite-based damage mapping following the 2006 Indonesia earthquake—How accurate was it? International Journal of Applied Earth Observation and Geoinformation 12, 466–476. https://doi.org/10.1016/j.jag.2010.07.004

Khoshelham, K., Oude Elberink, S., Sudan Xu, 2013. Segment-Based classification of damaged building roofs in aerial laser scanning data. IEEE Geoscience and Remote Sensing Letters 10, 1258–1262. https://doi.org/10.1109/LGRS.2013.2257676

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90. https://doi.org/10.1145/3065386

Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G., 2015. A convolutional neural network cascade for face detection. IEEE, pp. 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170

Lin, G., Shen, C., Hengel, A. van den, Reid, I., 2016. Efficient piecewise training of deep structured models for semantic segmentation. IEEE, pp. 3194–3203. https://doi.org/10.1109/CVPR.2016.348

Liu, W., Rabinovich, A., Culurciello, E., 2016. Parsenet: looking wider to see better, in: ICLR 2016. Presented at the ICLR 2016.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 55, 645–657. https://doi.org/10.1109/TGRS.2016.2612821

Marin, C., Bovolo, F., Bruzzone, L., 2015. Building change detection in multitemporal very high resolution SAR images. IEEE Transactions on Geoscience and Remote Sensing 53, 2664–2682. https://doi.org/10.1109/TGRS.2014.2363548

Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J., 2015. Multi-scale convolutional neura lnetworks for lung nodule classification, in: Ourselin, S., Alexander, D.C., Westin, C.-F., Cardoso, M.J. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 588–599. https://doi.org/10.1007/978-3-319-19992-4_46

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-Scale image recognition, in: ICLR 2015. pp. 1–13.

Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net, in: ICLR 2015.

Sui, H., Tu, J., Song, Z., Chen, G., Li, Q., 2014. A novel 3D building damage detection method using multiple overlapping UAV images. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-7, 173–179. https://doi.org/10.5194/isprsarchives-XL-7-173-2014

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Transactions on Medical Imaging 35, 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Tang, Y., Mohamed, A.R., 2012. Multiresolution deep belief networks, in: International Conference on Artificial Intelligence and Statistics. Presented at the International Conference on Artificial Intelligence and Statistics, Canary Islands, Spain.

Telgarsky, M., 2016. Benefits of depth in neural networks, in: 29th Annual Conference on Learning Theory. pp. 1–23.

Tu, J., Sui, H., Feng, W., Sun, K., Xu, C., Han, Q., 2017. Detecting building façade damage from oblique aerial images using local symmetry feature and the Gini Index. Remote Sensing Letters 8, 676–685. https://doi.org/10.1080/2150704X.2017.1312027

Ural, S., Hussain, E., Kim, K., Fu, C.-S., Shan, J., 2011. Building Extraction and Rubble Mapping for City Port-au-Prince Post-2010 Earthquake with GeoEye-1 Imagery and Lidar Data. Photogrammetric Engineering &

Remote Sensing 77, 1011–1023. https://doi.org/10.14358/PERS.77.10.1011

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2017. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS Journal of Photogrammetry and Remote Sensing. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Vetrivel, A., Gerke, M., Kerle, N., Vosselman, G., 2016b. Identification of structurally damaged areas in airborne oblique images using a Visual-Bag-of-Words approach. Remote Sensing 8, 231. https://doi.org/10.3390/rs8030231

Vetrivel, A., Kerle, N., Gerke, M., Nex, F., Vosselman, G., 2016a. Towards automated satellite image segmentation and classification for assessing disaster damage using data-specific features with incremental learning. Presented at the GEOBIA 2016, GEOBIA 2016, Enschede, The Netherlands. https://doi.org/10.3990/2.369

Vetrivel, A., Markus Gerke, Norman Kerle, George Vosselman, 2015. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. ISPRS Journal of Photogrammetry and Remote Sensing 105, 61–78. https://doi.org/10.1016/j.isprsjprs.2015.03.016

Wilson, C., Roelofs, R., Stern, M., Srebro, N., Recht, B., 2017. The marginal value of adaptive gradient methods in machine learning. arXiv:1705.08292.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions, in: ICLR 2016. Presented at the ICLR.

Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks, in: CVPR 2017. Presented at the CVPR 2017.

3 Multi-resolution feature fusion for the image classification of building damages[2]

[2] This chapter is based on the article:
Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. Remote Sens. 2018, 10, 1636

Abstract

Remote sensing images have long been preferred to perform building damage assessments. The recently proposed methods to extract damaged regions from remote sensing imagery rely on convolutional neural networks (CNN). The common approach is to train a CNN independently considering each of the different resolution levels (satellite, aerial, and terrestrial) in a binary classification approach. In this regard, an ever-growing amount of multi-resolution imagery are being collected, but the current approaches use one single resolution as their input. The use of up/down-sampled images for training has been reported as beneficial for the image classification accuracy both in the computer vision and remote sensing domains. However, it is still unclear if such multi-resolution information can also be captured from images with different spatial resolutions such as imagery of the satellite and airborne (from both manned and unmanned platforms) resolutions. In this chapter, three multi-resolution CNN feature fusion approaches are proposed and tested against two baseline (mono-resolution) methods to perform the image classification of building damages. Overall, the results show better accuracy and localization capabilities when fusing multi-resolution feature maps, specifically when these feature maps are merged and consider feature information from the intermediate layers of each of the resolution level networks. Nonetheless, these multi-resolution feature fusion approaches behaved differently considering each level of resolution. In the satellite and aerial (unmanned) cases, the improvements in the accuracy reached 2% while the accuracy improvements for the airborne (manned) case was marginal. The results were further confirmed by testing the approach for geographical transferability, in which the improvements between the baseline and multi-resolution experiments were overall maintained.

3.1 *Introduction*

The location of damaged buildings after a disastrous event is of utmost importance for several stages of the disaster management cycle. Manual inspection is not efficient since it takes a considerable amount of resources and time. Preventing the use of such inspections results in the early response phase of the disaster management cycle (United Nations, 2015). Over the last decade, remote sensing platforms have been increasingly used for the mapping of building damages. These platforms usually have a wide coverage, fast deployment, and high temporal frequency. Space, air, and ground platforms mounted with optical (Curtis and Fagan, 2013; Ishii et al., 2002; Vu et al., 2005), radar (Balz and Liao, 2010; Brunner et al., 2011), and laser (Armesto-González et al., 2010; Khoshelham et al., 2013) sensors have been used to collect data to perform automatic building damage assessment. Regardless of the platform and sensor used, several central difficulties persist, such as the subjectivity in the manual identification of hazard-induced damages from the

remote sensing data, and the fact that the damage evidenced by the exterior of a building might not be enough to infer the building's structural health. For this reason, most scientific contributions aim towards the extraction of damage evidence such as piles of rubble, debris, spalling, and cracks from remote sensing data in a reliable and automated manner.

Optical remote sensing images have been preferred to perform building damage assessments since these data are easier to understand when compared with other remote sensing data (Dell'Acqua and Gamba, 2012). Moreover, these images may allow for the generation of 3D models if captured with enough overlap. The 3D information can then be used to infer the geometrical deformations of the buildings. However, the time needed for the generation of such 3D information through dense image matching might hinder its use in the search and rescue phase because fast processing is mandatory in this phase.

Synoptic satellite imagery can cover regional to national extents and can be readily available after a disaster. The International Charter (IC) and the Copernicus Emergency Management Service (EMS) use synoptic optical data to assess building damage after a disastrous event. However, many signs of damage may not be identifiable using such data. Pancake collapses and damages along the façades might not be detectable due to the limited viewpoint of such platforms. Furthermore, its low resolution may introduce uncertainty in the satellite imagery damage mapping (Kerle and Hoffman, 2013), even when performed manually (Kerle, 2010; Saito et al., 2010).

To overcome these satellite imagery drawbacks, airborne images collected from manned aerial platforms have been considered in many events (Gerke and Kerle, 2011; Murtiyoso et al., 2014; Nex et al., 2014; Vetrivel et al., 2017). These images may not be as readily available as satellite data, but they can be captured at a higher resolution and such aerial platforms may also perform multi-view image captures. While the increase in the resolution aids in the disambiguation between damaged and non-damaged buildings, the oblique views enable the damage assessment of the façades (Gerke and Kerle, 2011). These advantages were also realized by the EMS, which recently started signing contracts with private companies to survey regions with aerial oblique imagery after a disaster ("CGR supplies aerial survey to JRC for emergency," n.d.), as it happened in the 2016 earthquakes in central Italy.

Unmanned aerial vehicles have been used to perform a more thorough damage assessment of a given scene. The high portability and higher resolution, when compared to manned platforms, have several benefits: they allow for a more detailed damage assessment (Vetrivel et al., 2017), which allows lower levels of damage such as cracks and smaller signs of spalling to be detected (Fernandez Galarreta et al., 2015), and they allow the UAV flights to focus only on specific areas of interest (Duarte et al., 2017).

Recent advances in the computer vision domain, namely, the use of convolutional neural networks (CNN) for image classification and segmentation (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2015), have also shown their potential in the remote sensing domain (Fu et al., 2017; Maggiori et al., 2017; Wei et al., 2017) and, more specifically, for the image classification of building damages such as debris or rubble piles (Vetrivel et al., 2016b, 2017). All these contributions use data with similar resolutions that are specifically acquired to train and test the developed networks. The use of multi-resolution data has improved the overall image classification and segmentation in many computer vision applications (Eigen and Fergus, 2015; Fu et al., 2017; Hu et al., 2015) and in remote sensing (Maggiori et al., 2017). However, multi-resolution images are generated artificially when the input images are up-sampled and down-sampled at several scales and then fused to obtain a final stronger classifier. While in computer vision, the resolution of a given image is considered as another inherent difficulty in the image classification task, in remote sensing, there are several resolution levels defined by the used platform and sensor, and these are usually considered independently for any image classification task.

A growing amount of image data have been collected by map producers using different sensors and with different resolutions, and their optimal use and integration would, therefore, represent an opportunity to positively impact scene classification. More specifically, a successful multi-resolution approach would make the image classification of building damages more flexible and not rely only on a given set of images from a given platform or sensor. This would be optimal since there often are not enough image samples of a given resolution level available to generate a strong CNN based classifier. The previous chapter focused on the satellite image classification of building damages (debris and rubble piles) whilst also considering image data from other (aerial) resolutions in its training. It was reported an improvement of nearly 4% in the satellite image classification of building damages by fusing the feature maps obtained from satellite and aerial resolutions. However, the chapter limited its investigation to satellite images, not considering the impact of the multi-resolution approach in the case of aerial (manned and unmanned) images.

The present chapter extends the previously reported work in the previous chapter, by thoroughly assessing the combined use of satellite and airborne (manned and unmanned) imagery for the image classification of the building damages (debris and rubble piles, as in Figure 13) of these same resolutions. This work focuses on the fusion of the feature maps coming from each of the resolutions. Specifically, the aim of the chapter is twofold:

- Assess the behavior of several feature fusion approaches by considering satellite and airborne (manned and unmanned) (Figure 13) feature

information, and compare them against two baseline experiments for the image classification of building damages;

- Assess the impact of multi-resolution fusion approaches in the model transferability for each of the considered resolution levels, where an image dataset from a different geographical region is only considered in the validation step.

The next section focuses on the related work of both image-based damage mapping and CNN feature map fusion. Section 3 presents the methodology followed to assess the use of multi-resolution imagery, where the used network is defined and the fusion approaches formalized. Section 4 deals with the experiments and results, followed by a discussion of the results (Section 5) and conclusions (Section 6).

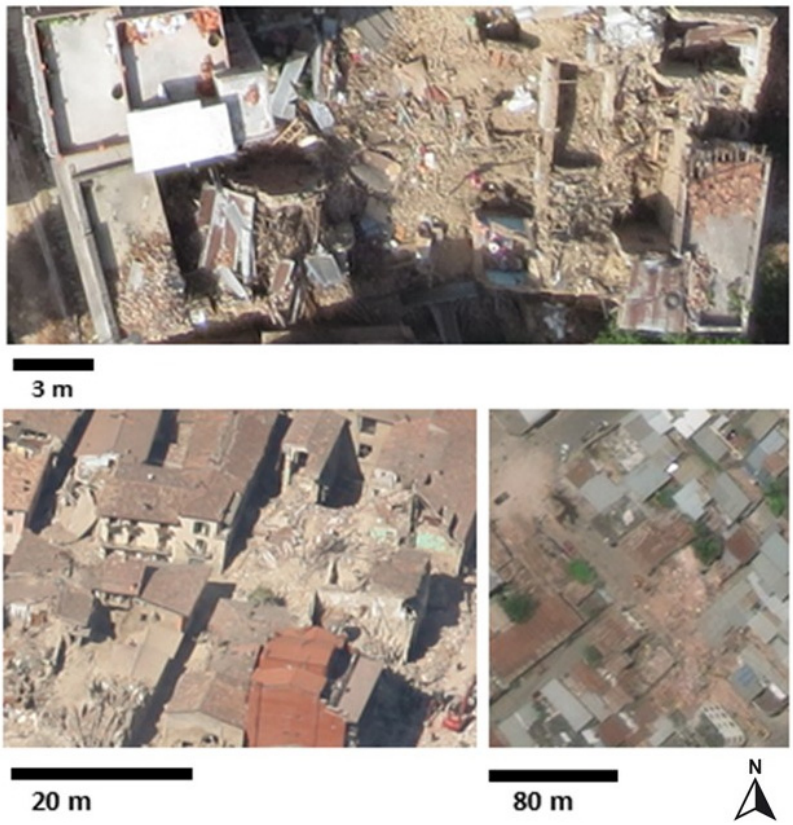

Figure 13. Examples of damaged and undamaged regions in remote sensing imagery. Nepal (top), aerial (unmanned). Italy (bottom left), aerial (manned). Ecuador (bottom right), satellite. These image examples also contain the type of damaged considered in this study: debris and rubble piles.

3.2 *Related Work*

3.2.1 Image-Based Damage Mapping

Various methods have been reported for the automatic image classification of building damages. These aim to relate the features extracted from the imagery with damage evidences. Such methods are usually closely related to the platform used for their acquisition, exploiting their intrinsic characteristics such as the viewing angle and resolution, among others. Regarding satellite imagery, texture features have been mostly used to map collapsed and partially collapsed buildings due to the coarse resolution and limited viewing angle of the platform. Features derived from the co-occurrence matrix have enabled the detection of partial and totally collapsed buildings from the IKONOS and QuickBird imagery (Vu et al., 2005). Multi-spectral image data from QuickBird, along with spatial relations formulated through a morphological scale-space approach have also been used to detect damaged buildings (Vu and Ban, 2010; Yamazaki et al., 2007). Another approach separated the satellite images into several classes; bricks and roof tiles were among them (Miura et al., 2007). The authors assumed that areas classified as bricks are most likely damaged areas.

The improvement of the image sensors coupled with the aerial platforms have not only increased the amount of detail present in aerial images but have also increased the complexity of the automation of damage detection procedures (Dong and Shan, 2013). Due to the high-resolution of the aerial imagery, object-based image analysis (OBIA) has started to be used to map damage (Li et al., 2011; Ma and Qin, 2012; Vetrivel et al., 2015) since objects in the scene are composed of a higher number of pixels. Instead of using the pixels directly, these approaches worked on the object level of an image composed of a set of pixels. In this way, the texture features were related not to a given pixel but to a set of pixels (Blaschke, 2010). Specifically, OBIA was used, among other techniques, to assess façades for damage (Fernandez Galarreta et al., 2015; Gerke and Kerle, 2011).

Overlapping aerial images can be used to generate 3D models through dense image matching, where 3D information can then be used to detect partial and totally collapsed buildings (Gerke and Kerle, 2011). Additionally, the use of fitted planes allows us to assess the geometrical homogeneity of such features and distinguish intact roofs from rubble piles. The 3D point cloud also allows for the direct extraction of the geometric deformations of building elements (Fernandez Galarreta et al., 2015), for the extraction of 3D features such as the histogram of the Z component of a normal vector (Vetrivel et al., 2017), and for the use of the aforementioned features alongside the CNN image features in a multiple-kernel learning approach (Gerke and Kerle, 2011; Vetrivel et al., 2017).

Videos recorded from aerial platforms can also be used to map damage. Features such as hue, saturation, brightness, edge intensity, predominant direction, variance, statistical features from the co-occurrence matrix, and 3D features have been derived from such video frames to distinguish damaged from non-damaged areas (Cusicanqui et al., 2018; Hasegawa et al., 2000; Mitomi et al., 2002).

Focusing on the learning approach from the texture features to build a robust classifier, Vetrivel et al. (2016a) used a bag-of-words approach and assumed that the damage evidence related to debris, spalling, and rubble piles shared the same local image features. The popularity of the CNN for image recognition tasks has successfully led to approaches that consider such networks for the image classification of building damage (satellite and aerial) (Duarte et al., 2018; Vetrivel et al., 2016b, 2016a).

Despite the recent advancements in computer vision, particularly in CNN, these works normally follow the traditional approach of having a completely separate CNN for each of the resolution levels for the image classification of building damages from remote sensing imagery (Vetrivel et al., 2017, 2016b). In this work, the use of a multi-resolution feature fusion approach is assessed.

3.2.2 CNN Feature Fusion Approaches in Remote Sensing

The increase in the amount of remote sensing data collected, be it from space, aerial, or terrestrial platforms, has allowed for the development of new methodologies which take advantage of the fusion of the different types of remote sensing data (Gomez-Chova et al., 2015). The combination of several streams of data in CNN architectures has also shown to improve the classification and segmentation results since each of the data modalities (3D, multi-spectral, RGB) contribute differently towards the recognition of a given object in the scene (Gomez-Chova et al., 2015; Paisitkriangkrai et al., 2015). While the presented overview focusses on CNN feature fusion approaches, there are also other approaches which do not rely on CNNs to perform data fusion (Hermosilla et al., 2011; Prince et al., 2017; Sohn and Dowman, 2007).

The fusion of 3D data from laser sensors or generated through dense image matching using images has been already addressed (Audebert et al., 2018, 2017; Paisitkriangkrai et al., 2015; Vetrivel et al., 2017). Liu et al. (2017) extracted handcrafted features from Lidar data alongside CNN features from the aerial images, fusing them in a higher order conditional random fields approach. Merging optical and Lidar data improved the semantic segmentation of 3D point clouds (Audebert et al., 2017) using a set of convolutions to merge both feature sets. The fusion of Lidar and multispectral imagery was also addressed (Audebert et al., 2018), in which the authors report the complementarity of such data in semantic segmentation. CNN and hand-crafted image features were concatenated to generate a stronger segmentation

network in the case of aerial images (Paisitkriangkrai et al., 2015). In the damage mapping domain, Vetrivel et al. (2017) merged both the CNN and 3D features (derived from a dense image-matching point cloud) in a multiple-kernel-learning approach for the image classification of building damages using airborne (manned and unmanned vehicles) images. The most relevant finding in this work was that the CNN features were so meaningful that, in some cases, the combined use of 3D information with CNN features only degraded the result, when compared to using only CNN features. The authors also found that CNNs still cannot optimally deal with the model geographical transferability in the specific case of the image classification of building damages because differences in urban morphology, architectural design, image capture settings, among others, may hinder this transferability.

The fusion of multi-resolution imagery coming from different resolution levels, such as satellite and airborne (manned and unmanned) imagery, had already been tested only for the specific case of satellite image classification of building damages (Duarte et al., 2018). The authors reported that it is more meaningful to perform a fusion of the feature maps coming from each of the resolutions than to have all the multi-resolution imagery share features in a single CNN. Nonetheless, the multi-resolution feature fusion approach was (1) not tested for the airborne (manned and unmanned) resolution levels and (2) not tested for the model transferability when a new region was only considered in the validation step.

3.3 *Methodology*

Three different CNN feature fusion approaches were used to assess the multi-resolution capabilities of CNN in performing the image classification of building damages. These multi-resolution experiments were compared with two baseline approaches. These baselines followed the traditional image classification pipeline using CNN, where each imagery resolution level was fed to a single network.

The network used in the experiments is presented in Section 3.1. This network exploited two main characteristics: residual connections and dilated convolutions (presented in the following paragraphs). The baseline experiments are presented in Section 3.2, while the feature fusion approaches are presented in Section 3.3.

A central aspect of a network capable of capturing multi-resolution information present in the images is its ability to capture spatial context. Yu and Koltun (2016) introduced the concept of dilated convolutions in CNN with the aim of capturing the context in image recognition tasks. Dilated convolutions are applied to a given input image using a kernel with defined gaps (Figure 14). Due to the gaps, the receptive field of the network is bigger, capturing more contextual information (Yu and Koltun, 2016). Moreover, the receptive field

size of the dilated convolutions also enables the capture of finer details since there is no need to perform an aggressive down-sampling of the feature maps throughout the network, better preserving the original spatial resolution (Yu et al., 2017). Looking at the specific task of building damage detection, the visual depiction of a collapsed building in a nadir aerial image patch may not appear in the form of a single rubble pile. Often, only smaller damage cues such as blown out debris or smaller portions of rubble are found in the vicinity of such collapsed buildings. Hence, by using dilated convolutions in this study, we aim to learn the relationship between damaged areas and their context, relating these among all the levels of resolution.

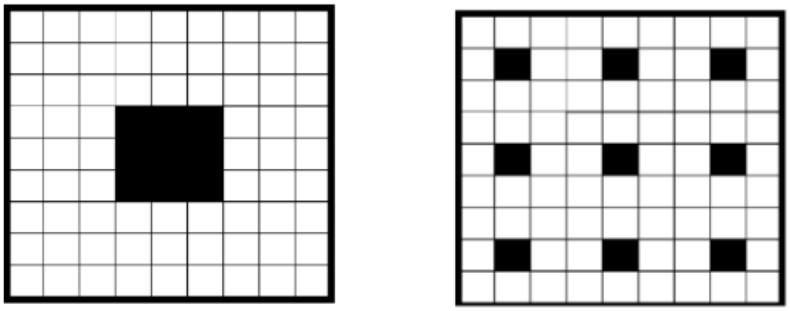

Figure 14. The scheme of (a) a 3 × 3 kernel with dilation 1, (b) a 3 × 3 kernel with dilation 3 (Duarte et al., 2018).

From the shallow *alexnet* (Krizhevsky et al., 2012), to the *VGG* (Simonyan and Zisserman, 2015), and the more recently proposed *resnet* (He et al., 2016), the depth of the proposed networks for image classification has increased. Unfortunately, the deeper the network, the harder it is to train (Simonyan and Zisserman, 2015). CNNs are usually built by the stacking of convolution layers, which allows a given network to learn from lower level features to higher levels of abstraction in a hierarchical setting. Nonetheless, a given layer l is only connected with the layers adjacent to it (i.e., layers $l-1$ and $l+1$). This assumption has shown to be not optimal since the information from earlier layers may be lost during backpropagation (He et al., 2016). Residual connections were then proposed (He et al., 2016), where the input of a given layer may be a summation of previous layers. These residual connections allow us to (1) have deeper networks while maintaining a low number of parameters and (2) to preserve the feature information across all layers (Figure 15) (He et al., 2016). The latter aspect is particularly important for a multi-resolution approach since a given feature may have a different degree of relevance for each of the considered levels of resolution. The preservation of this feature information is therefore critical when aggregating the feature maps generated using different resolution data.

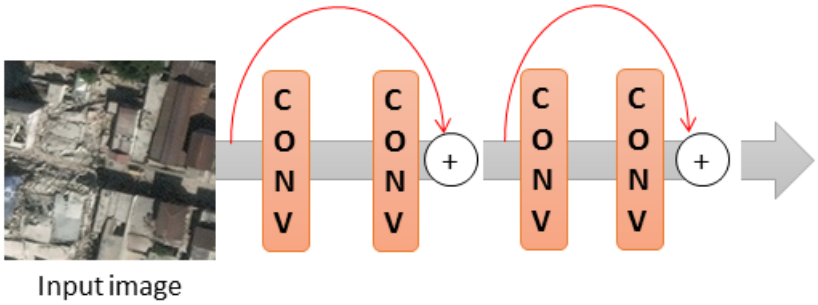

Figure 15. The scheme of a possible residual connection in a CNN. The grey arrows indicate a classical approach, while the red arrows on top show the new added residual connection (Duarte et al., 2018).

3.3.1 Basic Convolutional Set and Modules Definition

The main network configuration was built by considering two main modules: (1) the context module and (2) the resolution-specific module (Figure 16). This structure was inspired by the works of References [21,52,53]. The general idea regarding the use of these two modules was that while the dilated convolutions capture the wider context (context module), more local features may be lost in the dilation process, hence the use of the resolution-specific module (Hamaguchi et al., 2017; Yu and Koltun, 2016) with the decreasing dilation. In this way, the context is harnessed through the context module, while the resolution-specific module brings back the feature information related to a given resolution. The modules were built by stacking basic convolutional sets that were defined by convolution, batch normalization, and ReLU (rectified linear unit) (called CBR in Figure 16) (Ioffe and Szegedy, 2015). As depicted in Figure 16, a pair of these basic convolutional sets bridged by a residual connection formed the simplest component of the network, which were then used to build the indicated modules.

The context module was built by stacking 19 CBRs with an increasing number of filters and a dilation factor. For our tests, a lower number of CBRs would make the network weaker while deeper networks would give no improvements and slow the network runtime (increasing the risk of overfitting). The growing number of filters is commonly used in CNN approaches, following the general assumption that more filters are needed to represent more complex features (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). The increasing dilation factor in the context module is aimed at gradually capturing feature representations over a larger context area (Yu and Koltun, 2016). The red dots in Figure 16 indicate when a striding of 2, instead of 1, was applied. The striding reduced the size of the feature map (from the initial 224 × 224 px to the final 28 × 28 px) without performing max pooling. Larger striding has been shown to be beneficial when dilated convolutions are

considered (Yu et al., 2017). The kernel size was 3 × 3 (Springenberg et al., 2015) and only the first CBR block of the context module had a kernel size of 7 × 7 (Yu et al., 2017). The increase in the dilation factor can generate artifacts (aliasing effect) on the resulting feature maps due to the gaps introduced by the dilated kernels (Hamaguchi et al., 2017; Yu et al., 2017). To attenuate this drawback, the dilation increase in the context module was compensated in the resolution-specific module with a gradual reduction of the dilation value (Hamaguchi et al., 2017) and the removal of the residual connections from the basic CBR blocks (Yu et al., 2017). This also allowed us to recapture the more local features (Hamaguchi et al., 2017), which might have been lost due to the increasing dilations in the context module. For the classification part of the network, global average pooling followed by a convolution which maps the feature map size to the number of classes was applied (Long et al., 2015; Yu et al., 2017). Since this was a binary classification problem, a sigmoid function was used as the activation.

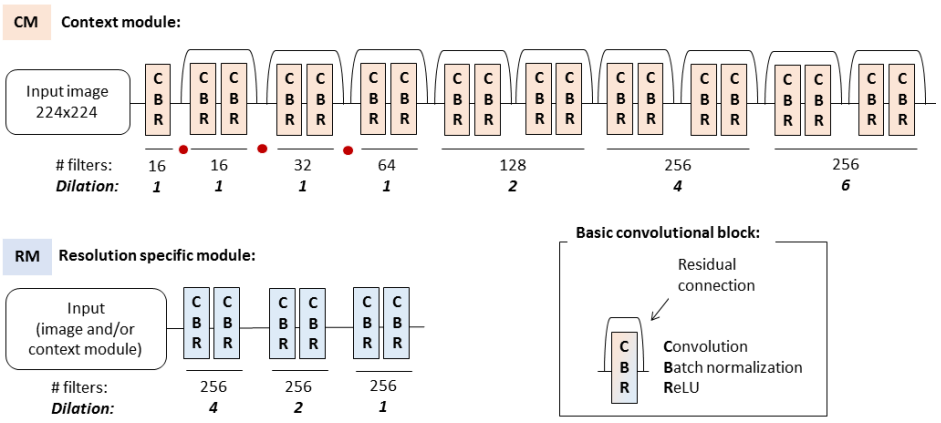

Figure 16. The basic convolution block is defined by convolution, batch-normalization, and ReLU (CBR). The CBR is used to define both the context and resolution-specific modules. It contains the number of filters used at each level of the modules and also the dilation factor. The red dot in the context module indicates when a striding of 2, instead of 1 was used.

3.3.2 Baseline Method

As already mentioned, the multi-resolution tests were compared against two baseline networks. These followed the traditional pipelines for the image classification of building damages [17,27].
In the first baseline network (Figure 17), the training samples of a single resolution (i.e., only airborne—manned or unmanned—or satellite) were fed into a network composed of the context and the resolution-specific module like in a single resolution approach. The second baseline (hereafter referred to as baseline_ft) used the same architecture as defined for the baseline (Figure 17).

It fed generic image samples of a given level of resolution (Table 5 and Table 6) into the context module, while the resolution-specific one was only fed with the damage domain image samples of that same level of resolution. Fine-tuning a network that used a generic image dataset for training may improve the image classification process (Maggiori et al., 2017), especially in cases with a low number of image samples for the specific classification problem (Tajbakhsh et al., 2016). The generic resolution-specific image samples were used to train a network considering two classes: built and non-built environments. Its weights were used as a starting point in the fine-tuning experiments for the specific case of the image classification of building damages. This led to two baseline tests for each resolution level (one trained from scratch and one fine-tuned on generic resolution-specific image samples).

3.3.3 Feature Fusion Methods

The multi-resolution feature fusion approaches used different combinations of the baseline modules and their computed features (Section 3.2). Three different approaches have been defined: MR_a, MR_b, and MR_c, as shown in Figure 5. The three types of fusion were inspired by previous studies in computer vision (Ngiam et al., 2011) and remote sensing (Audebert et al., 2018, 2017; Duarte et al., 2018; Gomez-Chova et al., 2015). In the presented implementation, the baselines were independently computed for each level of resolution without sharing the weights among them (Audebert et al., 2017). The used image samples have different resolutions and they were acquired in different locations: the multi-modal approaches (e.g., (Audebert et al., 2018)), dealing with heterogeneous data fusions (synchronized and in overlap), could not be directly adopted in this case as there was no correspondence between the areas captured by the different sensors. Moreover, in a disaster scenario, time is critical. Acquisitions with three different sensors (mounted on three different platforms) and resolutions would not be easily doable.

A fusion module (presented in Figure 5) was used in two of the fusion strategies, MR_b and MR_c, while MR_a followed the fusion approach used in Reference [30]. This fusion module aimed to learn from all the different feature representations, blending their heterogeneity (Audebert et al., 2018; Ngiam et al., 2011) through a set of convolutions. The objective behind the three different fusion approaches was to understand (i) which layers (and its features) were contributing more to the image classification of building damages in a certain resolution level and (ii) which was the best approach to fuse the different modules with multi-resolution information. The networks were then fine-tuned with the image data (X in Figure 5) of the resolution level of interest. For example, in MR_a, the features from the context modules of the three baseline networks were concatenated. Then, the resolution-specific module was fine-tuned with the image data X of a given resolution level (e.g., satellite imagery).

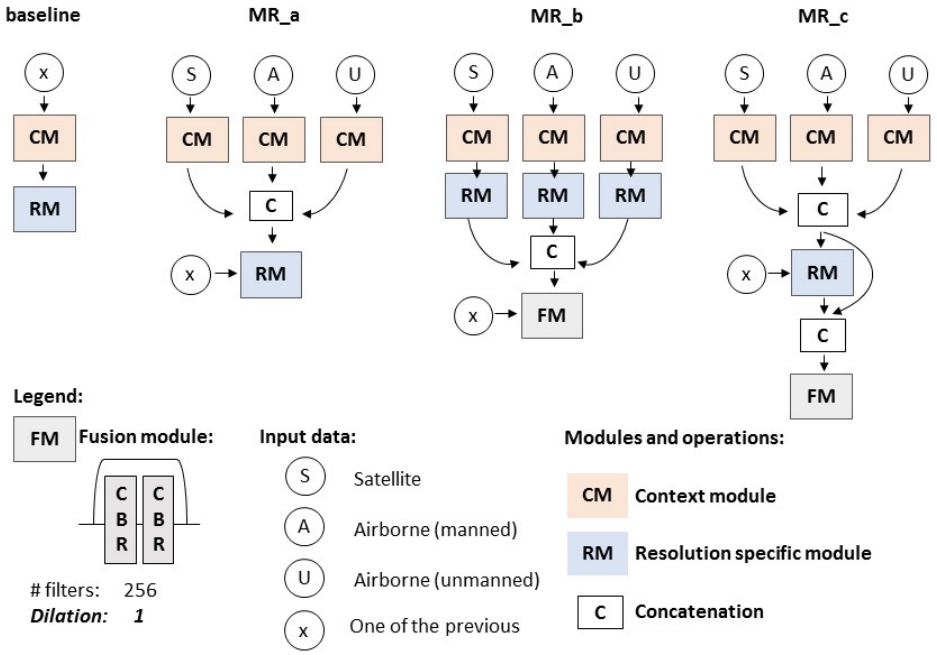

Figure 17. The baseline and multi-resolution feature fusion approaches (MR_a, MR_b, and MR_c). The fusion module is also defined.

The concatenation indicated in Figure 5 had as input the feature maps which had the same width and height, merging them along the channel dimension. Other merging approaches were tested such as summation, addition, and the averaging of the convolutional modules, however, they underperformed when compared to concatenation. In the bullet points below, each of the fusion approaches is defined in detail. Three fusions (MR_a, MR_b, and MR_c) were performed for each resolution level.

MR_a: in this fusion approach, the features of the context modules of each of the baseline experiments were concatenated. The resolution-specific module was then fine-tuned using the image data of a given resolution level (X, in Figure 5). This approach followed a general fusion approach already used in computer vision to merge the artificial multi-scale branches of a network (Eigen and Fergus, 2015; Li et al., 2015) or to fuse remote sensing image data (Boulch et al., 2017). Furthermore, this simple fusion approach has already been tested in another multi-resolution study (Duarte et al., 2018).

MR_b: in this fusion approach, the features of the context followed by the resolution-specific modules of the baseline experiments were concatenated. The fusion module considered as input the previous concatenation and it was fine-tuned using the image data of a given resolution level (*X*, in Figure 5). While only the context module of each resolution level was considered for the fusion in MR_a, MR_b considered the feature information of the resolution-

specific module. In this case, the fusion model aimed at blending all these heterogeneous feature maps and building the final classifier for each of the resolution levels separately (Figure 17). This fusion approach allows the use of traditional (i.e., mono resolution) pre-trained networks as only the last set of convolutions need to be run (i.e., fusion module).

MR_c: this approach builds on MR_a. However, in this case, the feature information from the concatenation of several context modules is maintained in a later stage of the fusion approach. This was performed by further concatenating this feature information with the output of the resolution-specific module that was fine-tuned with a given resolution image data (X in Figure 5). Like MR_b, the feature information coming from the context modules and resolution-specific module were blended using the fusion module.

3.4 *Experiments and Results*

The experiments, results, and used datasets are described in this section. The first set of experiments was performed to assess the classification results combining the multi-resolution data. In the second set of experiments, the model geographical transferability was assessed; i.e., when considering a new image dataset only for the validation (not used in training) of the networks.

3.4.1 Datasets and Training Samples

This subsection describes the datasets used in the experiments for each resolution level. It also describes the image sample generation from the raw images to image patches of a given resolution, which were then used in the experiments (Section 4.2). The data were divided into two main subsets: (a) a multi-resolution dataset formed by three sets of images corresponding to satellite and airborne (manned and unmanned) images containing damage image samples, and (b) three sets of generic resolution-specific image samples used in the fine-tuning baseline approach for the considered levels of resolution.

3.4.1.1 Damage Domain Image Samples for the Three Resolution Levels Considered

Most of the datasets depict real earthquake-induced building damages; however, there are also images of controlled demolitions (Table 1). The satellite images cover five different geographical locations in Italy, Ecuador, and Haiti. The satellite imagery was collected with WorldView-3 (Amatrice (Italy), Pescara del Tronto (Italy), and Portoviejo (Ecuador)) and GeoEye-1 (L'Aquila (Italy), Port-au-Prince (Haiti)). These data were pansharpened and have a variable resolution between 0.4 and 0.6 m. The airborne (manned platforms) images cover seven different geographic locations in Italy, Haiti,

and New Zealand. These sets of airborne data consist of nadir and oblique views. These were captured with the PentaView capture (Pictometry) and UltraCam Osprey (Microsoft) oblique imaging systems. Due to the oblique views, the ground sampling distance varies between 8 and 18 cm. These are usually captured with similar image capture specifications (flying height, overlap, etc.). The airborne (unmanned platforms) images cover nine locations in France, Italy, Haiti, Ecuador, Nepal, Germany, and China. These are composed of both the nadir and oblique views that were captured using both fixed wing and rotary wing aircraft mounted with consumer grade cameras. The ground sampling distance ranges from <1 cm up to 12 cm, where the image capture specifications (flying height, overlap, etc.) are related to the specific objective of each of the surveys, which changes significantly between the different datasets.

The image samples were derived from the set of images indicated before. First, the damaged and undamaged image regions were manually delineated, see Figure 18. A regular grid was then applied to each of the images and every cell that contained more than 40% of its area masked by the damage class was cropped from the image and used as an image sample for the damage class. The low value of 40% to consider a patch as damaged, aimed at forcing the networks to detect damage on an image patch even if it did not occupy the majority of the area of the said patch. This is motivated by practical reasons as an image patch should be considered damaged even if just a small area contains evidence of damage. On the other hand, a patch is considered intact only if no damage can be detected (Figure 18). The grid size varied according to the resolution: satellite = 80 × 80 px, airborne (manned vehicles) = 100 × 100 px, and airborne (unmanned) = 120 × 120 px (examples in Figure 19). The variable size of the image patches according to the resolution aimed to attenuate the captured extent by each of the resolution levels. The use of smaller patches also allowed us to increase the number of samples, compensating for the rare availability of these data.

Table 4. An overview of the location and quantity of the satellite and airborne image samples. The ++ locations indicate the controlled demolitions of buildings. Satellite used WorldView-3 GeoEye-1 imagery. Aerial manned used Vexcel and Pentaview systems while the Aerial unmanned used several commercial handheld cameras with varying characteristics.

Location	No. of Samples		Month/Year of Event
	Dam.	Not Dam.	
	Satellite		
L'Aquila (Italy)	115	108	4-2009
Port-au-Prince (Haiti)	701	681	1-2010
Portoviejo (Ecuador)	125	110	4-2016
Amatrice (Italy)	135	159	8-2016
Pesc. Tronto (Italy)	91	94	8- 2016
Total	1169	1152	
	Airborne (manned)		
L'Aquila (Italy)	242	235	4-2009
St Felice (Italy)	337	366	5-2012
Amatrice (Italy)	387	262	9-2016
Tempera (Italy)	151	260	4-2009
Port-au-Prince (slums) (Haiti)	409	329	1-2010
Port-au-Prince (Haiti)	302	335	1-2010
Onna (Italy)	293	265	2-2009
Christchurch (New Zealand)	603	649	2-2011
Total	2754	2701	
	Airborne (unmanned)		
L'Aquila (Italy)	103	99	4-2009
Wesel (Germany)	175	175	6-2016+
Portoviejo (Ecuador)	306	200	4-2016
Pesc. Tronto (Italy)	197	262	8-2016
Katmandu (Nepal)	388	288	4-2015
Taiwan (China)	257	479	2-2016
Gronau (Germany)	437	501	10-2013+
Mirabello (Italy)	412	246	5-2012
Lyon (France)	230	242	5-2017+
Total	2505	2692	

The number of image samples between the classes was approximately the same, while the number of image samples between the three different resolution levels was not balanced. The number of satellite image samples was two-fold lower when compared to the other two levels of resolution.

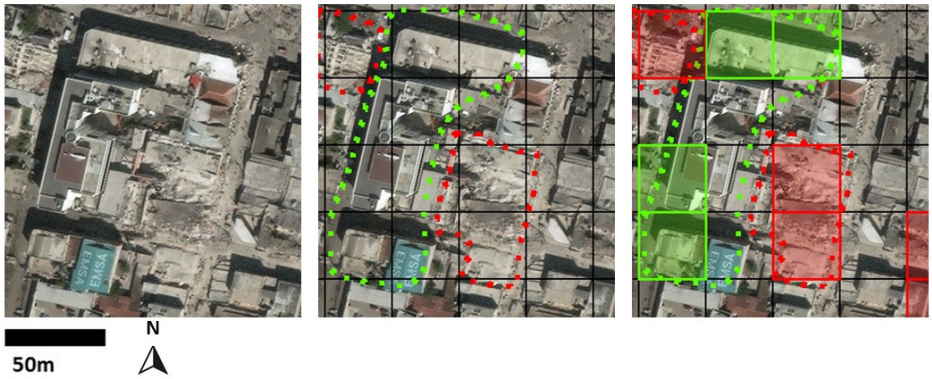

Figure 18. An example of the extracted samples considering a satellite image (GeoEye-1, Port-au-Prince, Haiti, 2010) on the left. The center image contains the grid for the satellite resolution level (80 × 80 px) where the damaged (red) and non-damaged (green) areas were manually digitized. The right patch indicates which squares of the grid are considered damaged and non-damaged after the selection process.

3.4.1.2 Generic Image Samples for the Three Levels of Resolution

Generic image samples for each of the levels of resolution are presented in this sub-section. These were used in one of the baseline approaches (baseline_ft).

The generic satellite image samples were taken from a freely available baseline dataset: NWPU-RESISC45 (Cheng et al., 2017). This baseline dataset contained 45 classes with 700 satellite image samples per class. From these, fourteen classes were selected and divided into two broader classes: built and non-built (Table 5).

To derive the generic image samples from the airborne images (manned and unmanned), the same sample extraction procedure used for the damage and non-damaged samples was adopted. In this case, the division was between the built and non-built environments, while the rest of the procedure was the same: a mask for the built and non-built environments was applied by considering a 60% threshold for each given class. This threshold was adopted to ensure that one of the two classes (the built and non-built environment classes) occupied the larger area of the image patch.

Table 5. The 14 classes of the benchmark dataset (NWPU-RESISC45) divided into the built and non-built classes. Each class contains 700 samples, with a total of 9800 image samples.

Built	**Non-Built**
Airport	Beach
Commercial area	Circular farmland
Dense residential	Desert
Freeway	Forest
Industrial area	Mountain
Medium residential	Rectangular farm
Sparse residential	Terrace

Table 6 shows the origin of the data for this generic image samples generation, the quantity of image samples and the considered camera for each location.

During the training of every network, data augmentation was performed (Table 4) since this was shown to decrease overfitting and improve the overall image classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). The used data augmentation consisted of random translations and rotations, image normalization, and the up-/down-sampling of the images (examples in Figure 20). Since we were dealing with oblique imagery in the airborne data, the performed flips were only horizontal and both the rotation value and the scale factor were low. Furthermore, light data augmentation is usually considered when batch normalization is used in a CNN since the network should be trained by focusing on less distorted images (Ioffe and Szegedy, 2015).

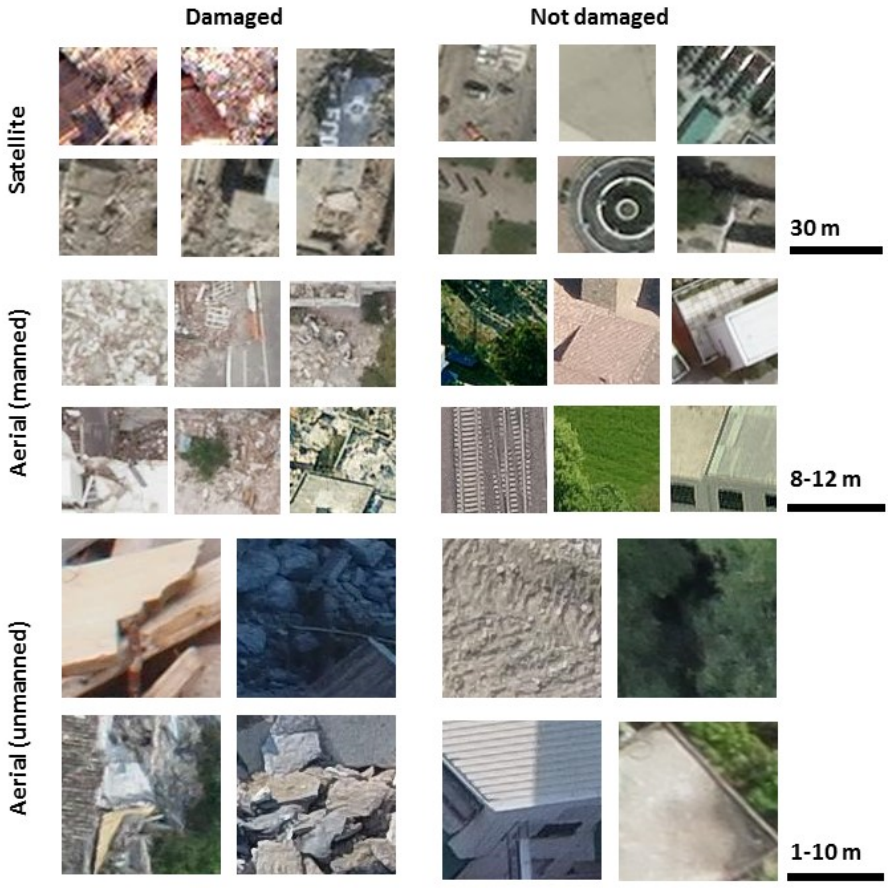

Figure 19. Examples of image samples derived from the procedure illustrated in Figure 6. These were used as the input for both the baseline and multi-resolution feature fusion experiments. (Left side) damaged samples; (Right side) non-damaged samples. From top to bottom: 2 rows of satellite, aerial (manned), and aerial (unmanned) image samples. The approximate scale is indicated for each resolution level.

Table 6. The generic airborne image samples used in one of the baselines. The * indicates that in the aerial (manned) case, three different locations from the Netherlands were considered. The system/sensor used are several handheld cameras for the unmanned aerial vehicles and PentaView and Vexcel imaging systems.

Location	Generic Airborne (Unmanned) Image Samples		Generic Airborne (Manned) Image Samples	
	Built	Non-Built	Built	Non-Built
Netherlands *	971	581	1758	878
France	697	690		
Germany	681	618	1110	1953
Italy	578	405		
Switzerland	107	688		
Total	3034	2982	2868	2831

Table 7. The data augmentation used: image normalization, the interval of the scale factor to be multiplied by the original size of the image sample, the rotation interval to be applied to the image samples, and the horizontal flip.

Data Augmentation	**Value**
Image normalization	1/255
Scale factor	[0.8,1.2]
Rotation	[−12,12] deg
Horizontal flip	true

The image samples were zero padded to fit in the 224 × 224 px input size, instead of being resized; this has been demonstrated to perform better (Vetrivel et al., 2016b) in the specific image classification of building damages using CNNs.

Two main sets of experiments were performed using the multi-resolution feature fusion approaches indicated in Figure 17: (1) general multi-resolution feature fusion experiments, where the training was performed using 70% of the image samples of each resolution and using the remaining 30% of the image samples for validation. This ratio was applied to each location separately. The training/validation data splits were performed randomly three times, enforcing the validation sets to contain different image samples on each data split; (2) model transferability, where the training of each of the multi-resolution feature fusion approaches was performed by considering all the locations except the one that was used for the validation. This experiment aimed at assessing the behavior of the approaches in a realistic scenario wherein the image data from a new event were classified without extracting any training samples from this location.

For both sets of experiments, the accuracy, recall, and precision were calculated for the validation image datasets described before and the following equations were considered:

$$accuracy = \frac{TP + FN}{\#\ validation\ samples} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

where, in Equations (1)–(3), TP are the true positives, FN are the false negatives, and FP are the false positives.

Figure 20. Several random data augmentation examples from an original aerial (unmanned) image sample with the scale, left.

3.4.2 Results

In this sub-section, the results of the multi-resolution fusion approaches are shown. The results are divided into two sub-sections for each of the resolution levels: the general multi-resolution fusion experiment and the model transferability experiment (using a dataset from a location not used in the training). To understand the behavior of the networks better, the activations from the last set of filters of the networks are visualized when classifying a new and unused image patch depicting a damaged scene. These activations show the per pixel probability of a pixel being damaged (white) or not damaged (black). Furthermore, in the model transferability sub-section, larger image patches were considered and classified with the best baseline and multi-resolution feature fusion approach.

3.4.2.1 Multi-Resolution Fusion Approaches

The achieved accuracies, recalls, and precisions for the baselines and for the different multi-resolution feature fusion approaches are presented in Table 8.

Table 8. The accuracy, recall, and precision results when considering the multi-resolution image data in the image classification of building damage of the given resolutions. Overall, the multi-resolution feature fusion approaches present the best results.

Network	**Satellite**			
	Accuracy(%)	Recall (%)	Precision (%)	Training Samples
baseline	87.7 ± 0.7	88.4 ± 0.9	87.4 ± 1.0	1602
baseline_ft	84.3 ± 0.8	84.1 ± 1.2	87.5 ± 1.8	11,402
MR_a	89.2 ± 1.0	87.0 ± 1.2	91.0 ± 1.3	8968
MR_b	89.3 ± 0.9	91.0 ± 0.9	86.5 ± 0.6	8968
MR_c	89.7 ± 0.9	93.1 ± 1.1	82.3 ± 1.6	8968
Network	**Airborne (Manned)**			
	Accuracy	Recall	Precision	Training Samples
baseline	91.1 ± 0.1	92.4 ± 1.5	91.1 ± 0.4	3736
baseline_ft	90.0 ± 0.4	89.8 ± 2.4	90.5 ± 0.3	9752
MR_a	91.4 ± 0.2	94.0 ± 0.6	88.0 ± 0.7	8968
MR_b	90.7 ± 0.4	91.9 ± 2.2	90.0 ± 1.2	8968
MR_c	91.4 ± 0.2	92.4 ± 0.7	89.4 ± 1.3	8968
Network	**Airborne (Unmanned)**			
	Accuracy	Recall	Precision	Training Samples
baseline	94.2 ± 1.0	93.1 ± 2.6	95.0 ± 0.7	3630
baseline_ft	91.3 ± 1.0	91.8 ± 2.0	89.9 ± 2.0	9329
MR_a	94.3 ± 0.7	94.1 ± 1.9	95.7 ± 1.9	8968
MR_b	95.3 ± 1.2	95.2 ± 0.7	95.3 ± 1.5	8968
MR_c	95.4 ± 0.6	95.5 ± 1.7	95.1 ± 1.2	8968

Considering the satellite resolution, the multi-resolution approaches improved the overall image classification of building damages when compared with the baselines by 2%. However, these also presented a slightly higher standard deviation between different runs. The MR_c presents the best results even though the improvement was marginal when compared with the other multi-resolution approaches. In comparison to the baseline experiment, the recall was higher in 2 of the 3 fusion approaches, while the precision was only higher in MR_a.

In the aerial (manned) case, the accuracy improvement was only marginal compared with the best performing baseline experiment. One of the multi-resolution approaches (MR_b) presented the worst results compared to the

baseline network. Baseline_ft was the experiment with the weakest performance as happened in the satellite case. MR_a had the highest recall and it also had the lower precision compared with the baseline experiment. MR_c increased the precision of the baseline test.

The airborne (unmanned) case also presented a marginal improvement using the proposed fusion approaches (MR_c and MR_b). Furthermore, in MR_c, the standard deviations of the experiments were lower. The baseline_ft was the experiment with the weakest performance. Overall, the best performing network regarding the classification accuracy was MR_c. This was further confirmed by the recall and precision values where all the fusion approaches had higher values for both the recall and precision than the baseline experiment.

The activations are shown in Figure 21. On the left, the input image patches are shown; on the right, the activations with the higher average activation value for each of the baseline and feature fusion approaches are shown. Overall, the multi-resolution fusion approaches presented better localization capabilities. These usually detected larger damaged areas than the baseline experiments. Namely, MR_c was the fusion approach with the better overall localization, even if it was noisier. The Figure 21 activations also present several striped patterns and gridding artifacts, where MR_b seems to be the network which better attenuates this issue.

3.4.2.2 Multi-Resolution Fusion Approaches' Impact on the Model Transferability

Table 6 shows the accuracies, recalls, and precisions of the multi-resolution and baseline approaches when using a single location in the validation which was not used in the training. In the satellite case, only the image data from Portoviejo were used as its validation data. In the airborne (manned) case, the Port-au-Prince image data were used as validation while in the airborne (unmanned) case, the Lyon image data were used for validation.

Table 9. The accuracy, recall and precision results when considering the multi-resolution feature fusion approaches for the model transferability. One of the locations for each of the resolutions is only used in the validation of the network: satellite = Portoviejo; aerial (manned) = Haiti; aerial (unmanned) = Lyon. Overall, the multi-resolution feature fusion approaches outperform the baseline experiments, where the baseline_ft present better results only in the aerial (manned) case.

Network	**Satellite (Portoviejo)**			
	Accuracy (%)	Recall (%)	Precision (%)	Training Samples
baseline	81.5	84	78	2160
baseline_ft	79.4	76	85	11,960
MR_a	81.5 ± 0.9	83.5 ± 0.1	83.5 ± 1.7	9526
MR_b	82.1 ± 0.6	77.7 ± 0.8	90.5 ± 1.5	9526
MR_c	83.4 ± 0.4	86.5 ± 0.9	82.9 ± 0.6	9526
Network	**Aerial (Manned, Port-au-Prince)**			
	Accuracy (%)	Recall (%)	Precision (%)	Training Samples
baseline	84.3	80.2	83.4	4406
baseline_ft	84.7	83.2	85.1	10,442
MR_a	81.9 ± 0.4	85.0 ± 0.3	78.6 ± 2.0	9638
MR_b	83.9 ± 0.4	80.3 ± 0.9	**84.1 ± 2.1**	9638
MR_c	84.2 ± 0.2	**85.0 ± 0.5**	80.0 ± 1.4	9638
Network	**Aerial (Unmanned, Lyon)**			
	Accuracy (%)	Recall (%)	Precision (%)	Training Samples
baseline	87.2	79.5	**95.1**	4711
baseline_ft	83.0	70.0	94.6	10,442
MR_a	85.7 ± 3.2	85.2 ± 3.6	90.0 ± 3.4	9943
MR_b	83.6 ± 2.1	86.2 ± 1.4	83.2 ± 3.3	9943
MR_c	**88.7 ± 1.7**	**89.6 ± 2.0**	82.4 ± 3.3	9943

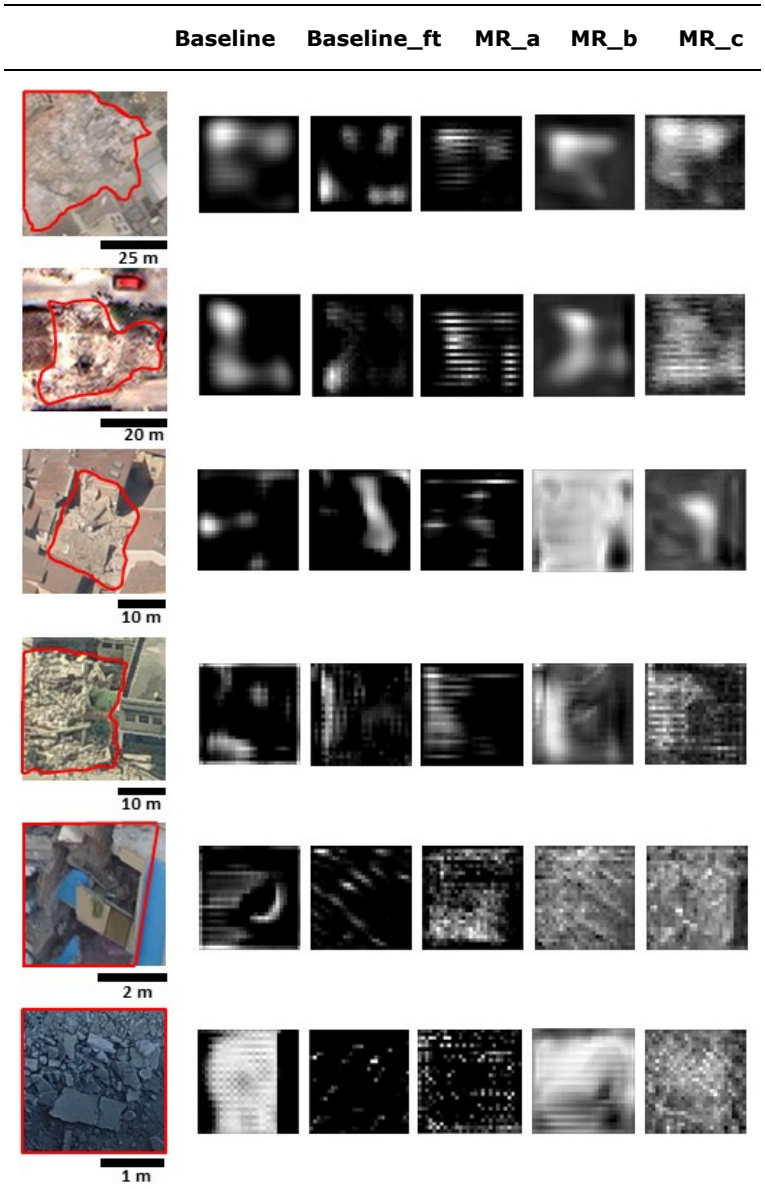

Figure 21. The image samples (left) and activations from the last set of feature maps (right) for each of the networks in the general multi-resolution feature fusion experiments. From top to bottom: 2 image samples of the satellite and aerial (manned and unmanned) resolutions. Overall, the multi-resolution feature fusion approaches have better localization capabilities than the baseline experiments.

Overall, the results followed the tendency of the previous experiments. The multi-resolution fusion approaches were the networks that performed better. Only in the aerial (manned) case was the baseline_ft accuracy superior to that

of the multi-resolution experiments. In the rest of the experiments, the baseline networks performed the worst.

In the airborne (unmanned) experiments, while the accuracy also increased with the MR_c feature fusion approach, the standard deviation was also considerably higher when compared with the rest of the experiments. Overall, the recall was higher in the fusion approaches, while the precision was lower when compared to the baseline experiments.

The activations are shown in Figure 22. On the left, the input image patches are shown; on the right, the activations with the highest average activation value per network are shown. Overall, the activations of the model transferability test presented the worst results when compared to the previous set of experiments. Striped patterns and gridding artifacts can also be noticed in this case. MR_b was the network which presented a lower amount of artifacts compared to the rest of the experiments. In the aerial (unmanned) case, the localization capability decreased drastically. Nonetheless, the multi-resolution experiments, in general, could better localize the damaged area.

In Figure 23 and Figure 25, larger image patches are shown for each of the locations considered for model transferability. These image patches were divided into smaller regions (80 × 80 px for the satellite, 100 × 100 px for the aerial manned, and 120 × 120 px for the aerial unmanned) and classified using the best performing baseline and multi-resolution feature fusion approaches (Table 9). The red overlay in these larger image patches indicates when a patch was classified as damaged (with a >0.5 probability of being damaged). The details (on the right) of these figures indicate the areas where differences between the baseline and the multi-resolution feature fusion methods were more significant. In these details, the probability of each of the smaller image patches being damaged is indicated.Figure 23 contains the image patch considered for the satellite level of resolution (Portoviejo). Besides correctly classifying 2 more patches as damaged, MR_c also increased the certainty of the already correctly classified patches in the baseline experiments. Nonetheless, none of the approaches was able to correctly classify the patch on the lower right corner of the larger image patch as damage.

Figure 24 shows a larger image patch for the aerial (manned) case (Port-au-Prince). The best performing networks were the baseline_ft and MR_c networks and the classification results are shown in the figure. In general, the results followed the accuracy assessment presented in Table 9. In this case, MR_c introduced more false positives (the details are on the right and on the bottom of the patch), even if it correctly classified more damaged patches.

Figure 25 shows a larger patch of the Lyon dataset classified with the benchmark and MR_c networks. In this case, the MR_c is clearly more generalizable. It reduced the false positives of the baseline approach and

correctly classified the patches that were not considered damaged by the baseline.

3.5 *Discussion*

The results show an improvement in the classification accuracy and the localization capabilities of a CNN for the image classification of building damages using the multi-resolution feature maps. However, each of the different feature fusion approaches behaved differently. The overall best multi-resolution feature fusion approach (MR_c) concatenates the feature maps from intermediate layers, confirming the need for preserving feature information from the intermediate layers at a later stage of the network (Eigen and Fergus, 2015; Maggiori et al., 2017).

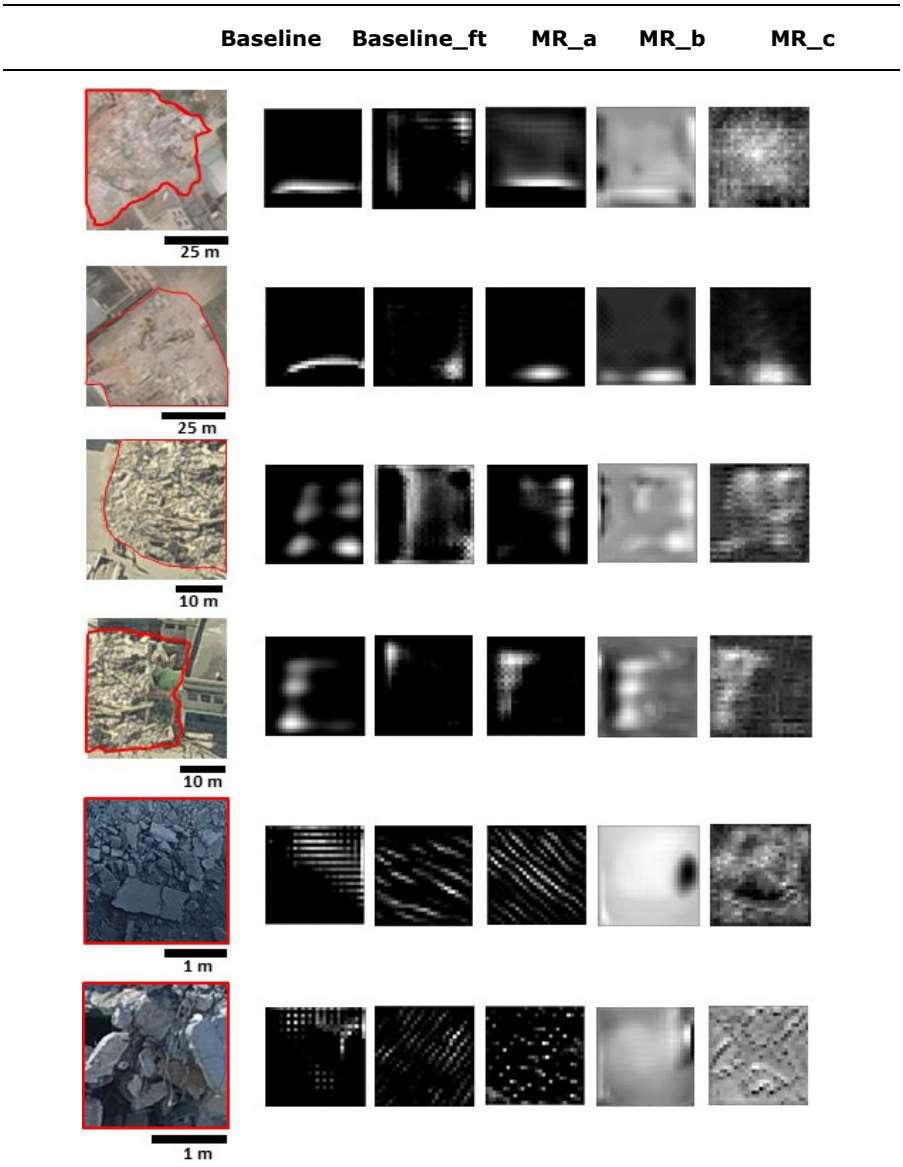

Figure 22. The image samples (left) and activations from the last set of the feature maps (right) for each of the networks in the model transferability experiments. From top to bottom: the 2 image samples of the satellite and aerial (manned and unmanned) resolutions.

This feature fusion approach also considers a fusion module (Figure 17) that is able to merge and blend the multi-resolution feature maps. Other feature fusion studies using small convolutional sets to merge audio and video features (Ngiam et al., 2011) or remote sensing multi-modal feature maps (Audebert et al., 2018; Liu et al., 2017; Paisitkriangkrai et al., 2015) have underlined the

same aspect. In general, the satellite and aerial (unmanned) resolutions were the ones which presented the most improvements when using multi-resolution feature fusion approaches. The aerial (unmanned) resolution also improved their image classification accuracy and localization capabilities (although marginally). In the aerial (manned) case, the resolution level had the least improvement with the multi-resolution feature fusion approach. This will be discussed in detail below.

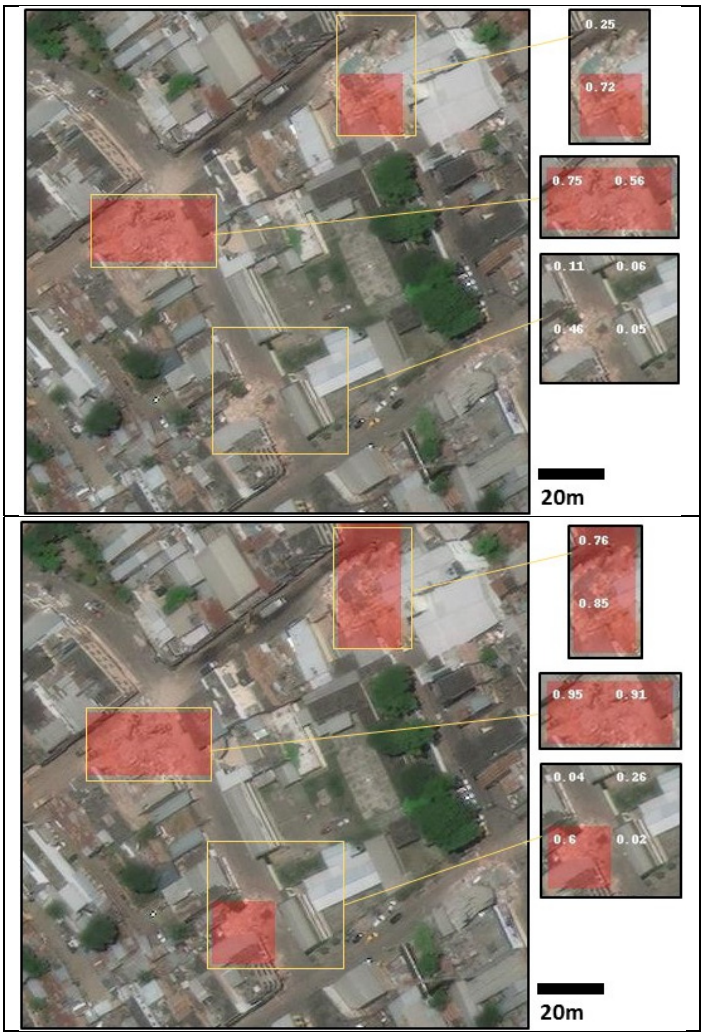

Figure 23. The large satellite image patch classified for damage using (top) the baseline and (bottom) the MR_c models on the Portoviejo dataset. The red overlay shows the image patches (80 × 80 px) considered as damaged (the probability of being damaged = >0.5). The right part with the details contains the probability of a given patch being damaged. The scale is relative to the large image patch on the left.

The model transferability experiments generally had a lower accuracy, indicating the need for in situ image acquisitions to get optimal classifiers, as shown in (Vetrivel et al., 2017). In the satellite case, both the precision and recall were higher in the multi-resolution feature fusion approaches, and the models captured fewer false positives and fewer false negatives. In the aerial (manned and unmanned) cases, the recall was higher and the precision was lower, reflecting that a higher number of image patches were correctly classified as damaged but more false positives were also present. In the aerial (manned) resolution tests, the multi-resolution feature fusion approaches had worse accuracies than the baselines. In this case, the best approach was to fine-tune a network which used generic aerial (manned) image samples during the training. In the aerial (manned) case, the image quality was better (high-end calibrated cameras), with more homogenous captures throughout different geographical regions. The aerial (unmanned) platform image captures were usually performed with a wide variety of compact grade cameras which presented a higher variability both in the sensor characteristics and in their image capture specifications. Consequently, there was a variable image quality compared to the aerial (manned) platforms.

The transferability tests of aerial (unmanned) imagery, contemporarily deal with geographical transferability aspects and also with very different image quality and image capture specifications. In such cases, the presented results indicate that the multi-resolution feature fusion approaches helped the model to be more generalizable than when using traditional mono-resolution methods.

The activations shown in the results are in agreement with the accuracy results. The multi-resolution feature fusion approaches presented better localization capabilities compared with the baseline experiments. Strike patterns and gridding artifacts can be seen in the activations. This could be due to the use of a dilated kernel in the presented convolutional modules, as indicated in (Hamaguchi et al., 2017; Yu et al., 2017).

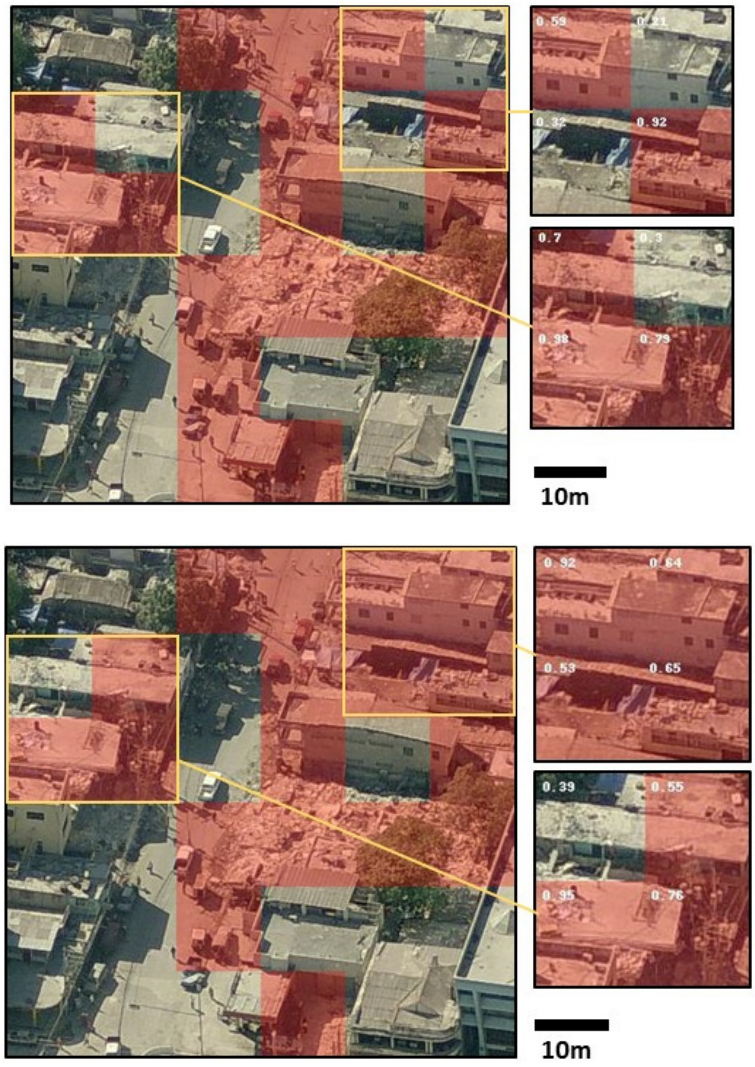

Figure 24. The large aerial (manned) image patch classified for damage using the (top) baseline_ft and (bottom) the MR_c models on the Port-au-Prince dataset. The red overlay shows the image patches (100 × 100 px) considered as damaged (the probability of being damaged = >0.5). The right part with the details contains the probability of a given patch being damaged. The legend is relative to the large image patch on the left.

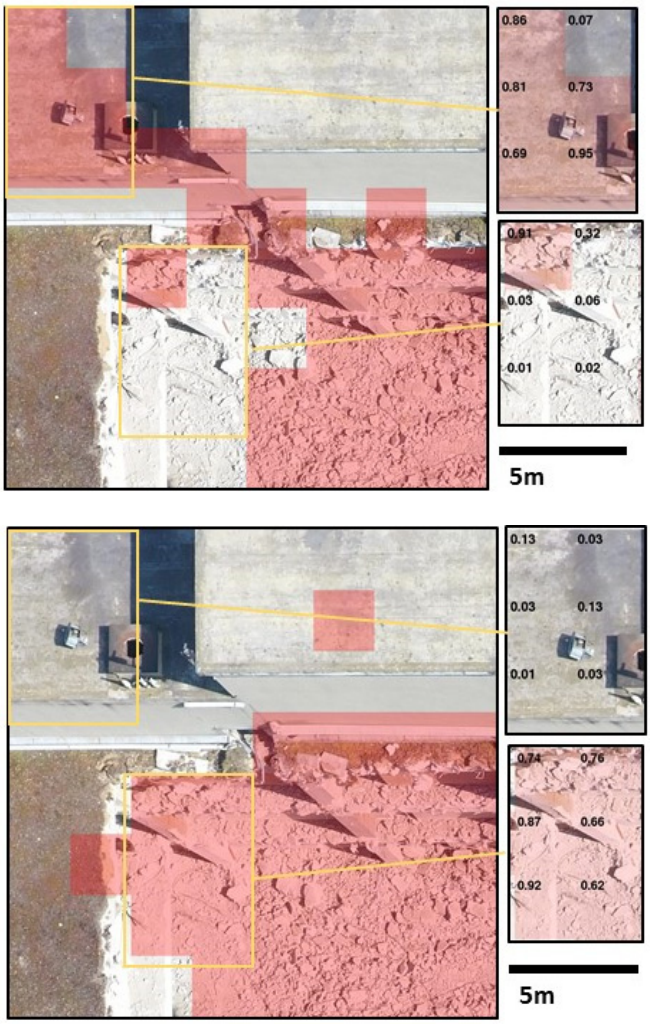

Figure 25. The large aerial (unmanned) image patch classified using (top) the baseline and (bottom) the MR_c models on the Lyon dataset. The red overlay shows the image patches (120 × 120 px) considered as damaged (the probability of being damaged = >0.5).

The right part of the figure shows the probability of each patch being damaged. The scale is relative to the large image patch on the left.

The large image patches shown in Figures 11 - 13 show that both the satellite and aerial (unmanned) resolution levels can benefit more from the multi-resolution feature fusion approach in comparison to the baseline experiments. Furthermore, the aerial (unmanned) multi-resolution feature fusion identifies only one of the patches as a false positive, while correctly classifying more damaged image patches.

The previous chapter, also on multi-resolution feature fusion, using both a baseline and a feature fusion approach similar to the MR_a, had better accuracies than the ones presented in this chapter, although both contributions reflect a general improvement. The differences in the two works is in the training data that were extracted from the same dataset but considering different images and different damage thresholds for the image patches labelling (40% in this chapter, 60% in the previous one). The different results confirm the difficulties and subjectivity inherent in the manual identification of building damages from any type of remote sensing imagery (Kerle, 2010; Saito et al., 2010). Moreover, it also indicates the sensibility of the damage detection with CNN according to the input used for training.

3.6 *Conclusions and Future Work*

This chapter assessed the combined use of multi-resolution remote sensing imagery coming from sensors mounted on different platforms within a CNN feature fusion approach to perform the image classification of building damages (rubble piles and debris). Both a context and a resolution-specific network module were defined by using dilated convolutions and residual connections. Subsequently, the feature information of these modules was fused using three different approaches. These were further compared against two baseline experiments.

Overall, the multi-resolution feature fusion approaches outperformed the traditional image classification of building damages, especially in the satellite and aerial (unmanned) cases. Two relevant aspects have been highlighted by the performed experiments on the multi-resolution feature fusion approaches: (1) the importance of the fusion module, as it allowed both MR_b and MR_c to outperform MR_a (2) the beneficial effect of considering the feature information from the intermediate layers of each of the resolution levels in the later stages of the network, as in MR_c.

These results were also confirmed in the classification of larger image patches in the satellite and aerial (unmanned) cases. Gridding artifacts and stripe patterns could be seen in the activations of the several fusion and baseline experiments due to the use of dilated kernels, however, in the multi-resolution feature fusion experiments, the activations were often more detailed than in the traditional approaches.

The model transferability experiments in the multi-resolution feature fusion approaches also improved the accuracy of the satellite and aerial (unmanned) imagery. On the contrary, fine-tuning a network by training it with generic aerial (manned) images was preferable in the aerial (manned) case. The different behavior in the aerial (manned) case could be explained by the use of images captured with high-end calibrated cameras and with more homogenous data capture settings. The characteristics of the aerial (manned)

resolution level contrasted with the aerial (unmanned) case, where the acquisition settings were more heterogeneous and a number of different sensors with a generally lower quality were used. In the aerial (manned) case, the model transferability to a new geographical region was, therefore, more related with the scene characteristics of that same region (e.g., urban morphology) and less related with the sensor or capture settings. In the aerial (unmanned) case, the higher variability of the image datasets allowed to better generalize the model.

The transferability test also indicated that the highest improvements of the multi-resolution approach were visible in the satellite resolution, with a substantial reduction of both false positives and false negatives. This was not the case in the aerial (unmanned) resolution level, where a higher number of false positives balanced the decrease in the number of false negatives. In a disaster scenario, the objective is to identify which buildings are damaged (hence, having potential victims). Therefore, it is preferable to lower the number of false negatives, maybe at the cost of a slight increase in false positives.

Despite the successful multi-resolution feature fusion approach for the image classification of building damages, there is no information regarding the individual contribution of each of the levels of resolution in the image classification task. Moreover, the presented results are mainly related to the overall accuracy and behavior of the multi-resolution feature fusion and baseline experiments. More research is needed to assess which signs of damage are better captured with this multi-resolution feature fusion approach, for each of the resolution levels. The focus of this work was on the fusion of the several multi-resolution feature maps. However, other networks can be assessed to perform the same task. In this regard, MR_b, for example, can be directly applied to pre-trained modules, where the last set of activations can be concatenated and posteriorly fed to the fusion module. In this case, there is no need to re-train a new network for a specific multi-resolution feature fusion approach. There is an ongoing increase in the amount of collected image data, where a multi-resolution approach could harness this vast amount of information and help build stronger classifiers for the image classification of building damages. Moreover, given the recent contributions focusing on online learning (Vetrivel et al., 2016b), the initial satellite images from a given disastrous event could be continuously refined with location-specific image samples that come from other resolutions. In such conditions, the use of a multi-resolution feature fusion approach would be optimal. This is especially relevant in an early post-disaster setting, where all these multi-resolution data would be captured independently with different sensors and at different stages of the disaster management cycle.

This multi-resolution feature fusion approach can also be assessed when considering other image classification problems with more classes. There is an

ever-growing amount of collected remote sensing imagery and taking advantage of this large quantity of data would be optimal.

3.7 *References of Chapter 3*

Armesto-González, J., Riveiro-Rodríguez, B., González-Aguilera, D., Rivas-Brea, M.T., 2010. Terrestrial laser scanning intensity data applied to damage detection for historical buildings. J. Archaeol. Sci. 37, 3037–3047. https://doi.org/10.1016/j.jas.2010.06.031

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011

Audebert, N., Le Saux, B., Lefèvre, S., 2017. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks, in: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (Eds.), Computer Vision – ACCV 2016. Springer International Publishing, Cham, pp. 180–196. https://doi.org/10.1007/978-3-319-54181-5_12

Balz, T., Liao, M., 2010. Building-damage detection using post-seismic high-resolution SAR satellite data. Int. J. Remote Sens. 31, 3369–3391. https://doi.org/10.1080/01431161003727671

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

Boulch, A., Saux, B.L., Audebert, N., 2017. Unstructured point cloud semantic labeling using deep segmentation networks, in: The Eurographics Association. https://doi.org/10.2312/3dor.20171047

Brunner, D., Schulz, K., Brehm, T., 2011. Building damage assessment in decimeter resolution SAR imagery: A future perspective, in: Joint Urban Remote Sensing Event. IEEE, pp. 217–220. https://doi.org/10.1109/JURSE.2011.5764759

CGR supplies aerial survey to JRC for emergency [WWW Document], n.d. . CGR Spa. URL http://www.cgrspa.com/news/cgr-fornira-il-jrc-con-immagini-aeree-per-le-emergenze/ (accessed 11.9.15).

Curtis, A., Fagan, W.F., 2013. Capturing damage assessment with a spatial video: an example of a building and street-scale analysis of tornado-related mortality in Joplin, Missouri, 2011. Ann. Assoc. Am. Geogr. 103, 1522–1538. https://doi.org/10.1080/00045608.2013.784098

Cusicanqui, J., Kerle, N., Nex, F., 2018. Usability of aerial video footage for 3D-scene reconstruction and structural damage assessment. Nat. Hazards Earth Syst. Sci. Discuss. 1–23. https://doi.org/10.5194/nhess-2017-409

Dell'Acqua, F., Gamba, P., 2012. Remote sensing and earthquake damage assessment: experiences, limits, and perspectives. Proc. IEEE 100, 2876–2890. https://doi.org/10.1109/JPROC.2012.2196404

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS J. Photogramm. Remote Sens. 84, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2018. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach, in: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 89–96. https://doi.org/10.5194/isprs-annals-IV-2-89-2018

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2017. Towards a more efficient detection of earthquake induced facade damages using oblique UAV imagery, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 93–100. https://doi.org/10.5194/isprs-archives-XLII-2-W6-93-2017

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: ICCV. IEEE, pp. 2650–2658. https://doi.org/10.1109/ICCV.2015.304

Fernandez Galarreta, J., Kerle, N., Gerke, M., 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Nat. Hazards Earth Syst. Sci. 15, 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sens. 9, 498. https://doi.org/10.3390/rs9050498

Gerke, M., Kerle, N., 2011. Automatic structural seismic damage assessment with airborne oblique Pictometry© imagery. Photogramm. Eng. Remote Sens. 77, 885–898. https://doi.org/10.14358/PERS.77.9.885

Gomez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G., 2015. Multimodal classification of remote sensing images: a review and future directions. Proc. IEEE 103, 1560–1584. https://doi.org/10.1109/JPROC.2015.2449668

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., Hikosaka, S., 2017. Effective use of dilated convolutions for segmenting small object instances in remote sensing images arXiv:1709.00179.

Hasegawa, H., Aoki, H., Yamazaki, F., Matsuoka, M., Sekimoto, I., 2000. Automated detection of damaged buildings using aerial HDTV images, in: IGARSS. IEEE, pp. 310–312. https://doi.org/10.1109/IGARSS.2000.860502

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR. https://doi.org/10.1109/CVPR.2016.90

Hermosilla, T., Ruiz, L.A., Recio, J.A., Estornell, J., 2011. Evaluation of Automatic Building Detection Approaches Combining High Resolution Images and LiDAR Data. Remote Sens. 3, 1188–1210. https://doi.org/10.3390/rs3061188

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7, 14680–14707. https://doi.org/10.3390/rs71114680

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift, in: 34th International Conference on Machine Learning. Sydney, Australia.

Ishii, M., Goto, T., Sugiyama, T., Saji, H., Abe, K., 2002. Detection of earthquake damaged areas from aerial photographs by using color and edge information, in: ACCV2002. Presented at the The 5th Asian Conference on Computer Vision, Melbourne, Australia.

Kerle, N., 2010. Satellite-based damage mapping following the 2006 Indonesia earthquake—How accurate was it? Int. J. Appl. Earth Obs. Geoinformation 12, 466–476. https://doi.org/10.1016/j.jag.2010.07.004

Kerle, N., Hoffman, R.R., 2013. Collaborative damage mapping for emergency response: the role of Cognitive Systems Engineering. Nat. Hazards Earth Syst. Sci. 13, 97–113. https://doi.org/10.5194/nhess-13-97-2013

Khoshelham, K., Oude Elberink, S., Sudan Xu, 2013. Segment-based classification of damaged building roofs in aerial laser scanning data. IEEE Geosci. Remote Sens. Lett. 10, 1258–1262. https://doi.org/10.1109/LGRS.2013.2257676

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1907–1105.

Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G., 2015. A convolutional neural network cascade for face detection, in: CVPR. IEEE, pp. 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170

Li, X., Yang, W., Ao, T., Li, H., Chen, W., 2011. An improved approach of information extraction for earthquake-damaged buildings using high-resolution imagery. J. Earthq. Tsunami 05, 389–399. https://doi.org/10.1142/S1793431111001157

Liu, Y., Piramanayagam, S., Monteiro, S.T., Saber, E., 2017. Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs, in: CVPR. IEEE, pp. 1561–1570. https://doi.org/10.1109/CVPRW.2017.200

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: CVPR. IEEE, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Ma, J., Qin, S., 2012. Automatic depicting algorithm of earthquake collapsed buildings with airborne high resolution image, in: International Geoscience and Remote Sensing Symposium. IEEE, pp. 939–942. https://doi.org/10.1109/IGARSS.2012.6351400

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans.

Geosci. Remote Sens. 55, 645–657. https://doi.org/10.1109/TGRS.2016.2612821

Mitomi, H., Matsuoka, M., Yamazaki, F., 2002. Application of automated damage detection of buildings due to earthquakes by panchromatic television images, in: The 7th US National Conference on Earthquake Engineering. Presented at the The 7th US National Conference on Earthquake Engineering.

Miura, H., Yamazaki, F., Matsuoka, M., 2007. Identification of damaged areas due to the 2006 Central Java, Indonesia earthquake using satellite optical images, in: Urban Remote Sensing Joint Event. IEEE, pp. 1–5. https://doi.org/10.1109/URS.2007.371867

Murtiyoso, A., Remondino, F., Rupnik, E., Nex, F., Grussenmeyer, P., 2014. Oblique aerial photography tool for building inspection and damage assessment, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 309–313. https://doi.org/10.5194/isprsarchives-XL-1-309-2014

Nex, F., Rupnik, E., Toschi, I., Remondino, F., 2014. Automated processing of high resolution airborne images for earthquake damage assessment, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 315–321. https://doi.org/10.5194/isprsarchives-XL-1-315-2014

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A., 2011. Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning.

Paisitkriangkrai, S., Sherrah, J., Janney, P., Van-Den Hengel, A., 2015. Effective semantic pixel labelling with convolutional networks and Conditional Random Fields, in: CVPR. IEEE, pp. 36–43. https://doi.org/10.1109/CVPRW.2015.7301381

Prince, D., Sidike, P., Essa, A., Asari, V., 2017. Multifeature fusion for automatic building change detection in wide-area imagery. J. Appl. Remote Sens. 11, 026040. https://doi.org/10.1117/1.JRS.11.026040

Saito, K., Spence, R., Booth, E., Madabhushi, G., Eguchi, R., Gill, S., 2010. Damage assessment of Port -au-Prince using Pictometry, in: 8th International Conference on Remote Sensing for Disaster Response. Tokyo Institute of Technology.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: ICLR. pp. 1–13.

Sohn, G., Dowman, I., 2007. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction. ISPRS J. Photogramm. Remote Sens. 62, 43–63. https://doi.org/10.1016/j.isprsjprs.2007.01.001

Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: the all convolutional net, in: ICLR.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35, 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

United Nations, 2015. INSARAG guidelines, volume II: preparedness and response, manual B: operations.

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2017. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Vetrivel, A., Gerke, M., Kerle, N., Vosselman, G., 2016a. Identification of structurally damaged areas in airborne oblique images using a Visual-Bag-of-Words approach. Remote Sens. 8, 231. https://doi.org/10.3390/rs8030231

Vetrivel, A., Gerke, M., Kerle, N., Vosselman, G., 2015. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. ISPRS J. Photogramm. Remote Sens. 105, 61–78. https://doi.org/10.1016/j.isprsjprs.2015.03.016

Vetrivel, A., Kerle, N., Gerke, M., Nex, F., Vosselman, G., 2016b. Towards automated satellite image segmentation and classification for assessing disaster damage using data-specific features with incremental learning, in: GEOBIA 2016. GEOBIA 2016, Enschede, The Netherlands. https://doi.org/10.3990/2.369

Vu, T.T., Ban, Y., 2010. Context-based mapping of damaged buildings from high-resolution optical satellite images. Int. J. Remote Sens. 31, 3411–3425. https://doi.org/10.1080/01431161003727697

Vu, T.T., Matsuoka, M., Yamazaki, F., 2005. Detection and Animation of Damage Using Very High-Resolution Satellite Data Following the 2003 Bam, Iran, Earthquake. Earthq. Spectra 21, 319–327. https://doi.org/10.1193/1.2101127

Wei, Y., Wang, Z., Xu, M., 2017. Road structure refined CNN for road extraction in aerial images. IEEE Geosci. Remote Sens. Lett. 14, 709–713. https://doi.org/10.1109/LGRS.2017.2672734

Yamazaki, F., Vu, T.T., Matsuoka, M., 2007. Context-based detection of post-disaster damaged buildings in urban areas from satellite images, in: Urban Remote Sensing Joint Event. IEEE, pp. 1–5. https://doi.org/10.1109/URS.2007.371869

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions, in: ICLR.

Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks, in: CVPR.

4 Towards a more efficient detection of earthquake induced façade damages using oblique UAV imagery[3]

[3] This chapter is based on the article:
Duarte, D., Nex, F., Kerle, N., and Vosselman, G.: Towards a more efficient detection of earthquake induced façade damages using oblique UAV imagery, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W6, 93-100, https://doi.org/10.5194/isprs-archives-XLII-2-W6-93-2017, 2017.

Abstract

Urban search and rescue (USaR) teams require a fast and thorough building damage assessment, to focus their rescue efforts accordingly. Unmanned aerial vehicles (UAV) are able to capture relevant data in a short time frame and survey otherwise inaccessible areas after a disaster, and have thus been identified as useful when coupled with RGB cameras for façade damage detection. Existing literature focuses on the extraction of 3D and/or image features as cues for damage. However, little attention has been given to the efficiency of the proposed methods which hinders its use in an urban search and rescue context. The framework proposed in this chapter aims at a more efficient façade damage detection using UAV multi-view imagery. This was achieved directing all damage classification computations only to the image regions containing the façades, hence discarding the irrelevant areas of the acquired images and consequently reducing the time needed for such task. To accomplish this, a three-step approach is proposed: i) building extraction from the sparse point cloud computed from the nadir images collected in an initial flight; ii) use of the latter as proxy for façade location in the oblique images captured in subsequent flights, and iii) selection of the façade image regions to be fed to a damage classification routine. The results show that the proposed framework successfully reduces the extracted façade image regions to be assessed for damage 6 fold, hence increasing the efficiency of subsequent damage detection routines. The framework was tested on a set of UAV multi-view images over a neighborhood of the city of L'Aquila, Italy, affected in 2009 by an earthquake.

4.1 *Introduction and related work*

Early post-disaster efforts, in particular the delineation and optimization of urban search and rescue (USaR) deployment, require fully automated, fast and detailed building damage assessment. This detailed damage information aids in the identification of viable rescue sites and is commonly performed by an USaR mobile team (United Nations 2015). However, in a hazard event such as an earthquake, ground observations have several limitations: limited access/points of view, procedure requiring a substantial amount of time and the need of sufficient USaR personnel.

Remote sensing has been recognized as a critical aid in building damage assessment (Dong and Shan 2013). Optical (Dell'Acqua and Polli 2011; Vetrivel et al. 2017), radar (Gokon et al. 2015; Marin, Bovolo, and Bruzzone 2015) or laser instruments (Armesto-González et al. 2010; Khoshelham, Oude Elberink, and Sudan Xu 2013) have already been used successfully in building damage detection. These, mounted on aerial platforms may acquire data in a short time interval and allow the automatization of the damage detection procedures (Dell'Acqua and Gamba 2012).

In particular, aerial images have been demonstrated to be suited for building damage assessment (Dong and Shan 2013; Vetrivel, Gerke, et al. 2016). The use of overlapping images allows for the computation of 3D point clouds, adding geometric information to the radiometric content of the images. While the point clouds are usually used to detect damages in the form of geometrical deformations (e.g. collapsed building), the images are used to detect damage evidences which may not be clearly represented in the point cloud (e.g. cracks or spalling) (Fernandez Galarreta, Kerle, and Gerke 2015; Sui et al. 2014; Vetrivel, Duarte, et al. 2016).

Nadir aerial imagery readily depicts totally collapsed buildings or damaged roofs (Ma and Qin 2012). However, nadir imagery is physically constrained by its capture geometry and cannot directly observe the façades. Even a pancake collapse of a building or a partially collapsed façade with an otherwise intact roof cannot be directly identified.

To overcome this limitation, airborne multi-view images started to be exploited for building damage assessment. With this capture geometry it is possible to survey directly the façades, and consequently, assess them for damage evidences (Gerke and Kerle 2011). Nonetheless, unmanned aerial vehicles (UAV) with their fast data acquisition, centimetre resolution, high revisit capability, low cost and possibility of surveying otherwise inaccessible or dangerous areas, seem to be the fit-for-purpose platform for USaR building damage assessment.

Similar to the airborne, the UAV multi-view images are usually collected with enough overlap to derive a 3D point cloud through the computationally expensive dense image matching (DIM). This allows to assess geometrical deformations through the extraction of 3D features (Fernandez Galarreta et al. 2015; Vetrivel et al. 2017). Assuming that a given façade is still standing or is only partially collapsed, the image information becomes critical to identify damage evidences that may not be translated into any deformation in an image-derived point cloud (Fernandez Galarreta et al. 2015). The relevance of the images for damage detection was also pointed out by Vetrivel et al. (2017). The authors indicated the negligible increase in accuracy when using 3D features and convolutional neural network (CNN) features in a multiple-kernel-learning approach, instead of the CNN features alone. This approach reached an average damage detection accuracy of 85%, solely using CNN features derived from labelled image samples from UAV datasets, not containing samples from the dataset being analysed for damage.

When the 3D information is not generated, the time needed for the damage detection part is reduced. However, the processing time is still lengthy, due to the high amount of images that are usually collected in such UAV-multi-view surveys.

Procedures like the simple linear iterative clustering (SLIC) (Achanta et al. 2012) segmentation are often used as starting point for current object-based or damage classification procedures, as in the CNN approach indicated earlier. These are applied to every image of a given dataset, which is not efficient. The temporal inefficiency is not a problem in many applications but limits the use of such methods in the USaR context.

The objective of this contribution is to propose a more efficient approach for a thorough façade damage detection using UAV multi-view imagery. Specifically, the aim is to avoid the computationally expensive procedures, and to direct all damage classification computations only to the images and image portions containing the façades, hence discarding the irrelevant areas of the captured UAV images. To accomplish this, a three-step approach is proposed, taking advantage of the rapid data acquisition and ready revisiting capabilities of the UAV (see Figure 27): i) extract the building's roof outline from the sparse point cloud generated from nadir images alone; ii) use the latter as a proxy for façade location in the oblique images, using the raw image orientation information of the UAV, and iii) damage detection only on relevant patches of the extracted façade image patch using the CNN as in Vetrivel et al. (2017). More details regarding the method are given in section 3.

The remainder of the chapter contains in section 2, a description of the data used in the experiment. Section 4, contains the results, followed by discussion and conclusion, in sections 5 and 6, respectively.

4.2 *Data*

The proposed approach was tested on a set of UAV multi-view images, captured using a Sony ILCE-6000 mounted on an Aibot X6 hexacopter. It comprises a subset of 281 nadir images, and four subsets of oblique images (891 images in total, one set for each cardinal direction). These were captured using a flying height of approximately 100 m with 70-80% forward overlap and 60-70% side lap. The average ground sampling distance is ~0.04m.

The captured images depict the damage caused by the M5.9 April 6th 2009 earthquake in L'Aquila, Italy. These were acquired over a city block of approximately10 ha. The scene contains partial collapses and buildings with smaller signs of damage (e.g. cracks and spalling). In spite of the image capture only being performed in 2016, the area of the city covered was abandoned and still contains the damage evidences from the 2009 earthquake, with only very limited reconstruction taking place. Due to the time interval between event and capture, and since the area is still largely untouched since the earthquake, it contains several areas with high vegetation. Hence, many of the façades are not visible, even in the oblique images (see Figure 26), making this dataset more challenging for research purposes.

Figure 26 Three examples of vegetation occlusion in the UAV multi-view L'Aquila dataset

4.3 *Method*

The central idea behind the targeted efficiency increase in façade damage mapping, is to reduce not only the number of images that are used in a façade damage detection routine, considering a conventional grid flight; but also to reduce the area of the oblique images to be fed for damage classification. In a first stage the façades are defined. This façade location allows to select only the oblique images that contain a given façade. Moreover, knowing the façade location also enables the identification of the oblique image patch corresponding to a given façade. Only this patch is then fed to the damage detection step. The second core idea regarding this method is to avoid that the whole façade image patch is fed to the damage assessment phase. The façade image patch is divided into equilateral patches of a given size, where only patches with early evidence of damage are fed to the damage classification step, which will use a pre-trained CNN, more details in section 3.3.

The approach can be divided in three main steps as presented in Figure 27. The initial step is to detect the buildings, that will be used as proxies for the presence of façades. The second step is to use the façade locations to extract the façade patch from the oblique images. The last step refers to the façade damage detection on the previously extracted façades.

The first step of the method is to locate the façades, as shown in Figure 28. Considering that every façade is connected to a building roof, this need to be located and a building hypothesis formulated, to subsequently define the façades. Usually the DIM point cloud is used as the main source of information to perform the building extraction phase. This is due to the general assumption

that building roofs are elevated (above ground) objects composed by planar surfaces.

Since one of the aims of the proposed approach is to avoid the computationally expensive DIM, it is hypothesized that to detect the building's roof, the (sparse) tie point cloud suffices. A conventional UAV nadir flight generates a large amount of images, and it is expected that the sparse point cloud is dense enough to derive building information. To reduce the number of outliers only tie points present in at least three images are considered.

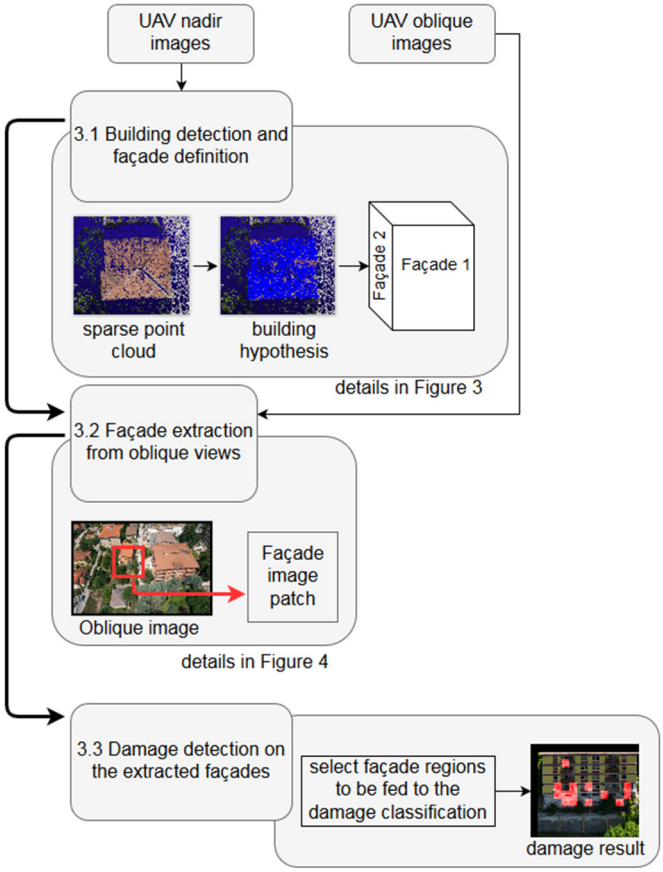

Figure 27 Overview of the method - divided into the three main components

4.3.1 Building detection and façade extraction

The sparse point cloud is generated using *Pix4D*, which also generates the internal camera parameters and the updated orientation of the images. In a first step, a distinction is needed from *on* and *off* ground points, to identify the elevated objects present in the scene. This is achieved recurring to *LAStools*

software package which uses the method proposed by Axelsson (2000). Due to the common heterogeneity of sparse point clouds, since these rely on the texture present in the scene to compute the point correspondences, isolated points are removed with *lasnoise*. This is performed to avoid the inclusion of these isolated points in the building detection phase. With the isolated points removed, the following step is to differentiate between *on* and *off* ground points, using *lasground*. This further allows to obtain a normalized height surface by differencing each of the *off ground* points by its closest *on groud* point.

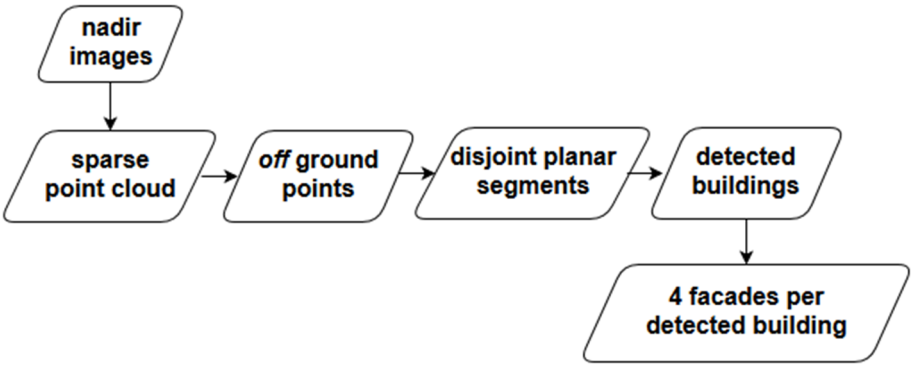

Figure 28 Building extraction and facade definition flowchart

1) Building detection from the *off* ground points: the *off ground* points of the sparse point cloud are segmented into disjoint planar segments as described in Vosselman (2012). An initial set of 10 points is used to estimate the plane equation and initialize the region growing algorithm. An initial distance threshold of 0.3 m is used to determine these points. New points are added considering a radius of 2 m to define the local neighbourhood: only those that have a distance from the plane lower than 1 m are added. These adopted parameters are intentionally lax in order to address the low and heterogeneous point density of some building roofs. Since there still may exist points on vertical elements of building roofs, segments with a slope greater than 70% are discarded.The resulting segments are then merged into building regions using a connected component analysis.

2) Façades per detected building: the points belonging to a given building region are initially projected into the *xy* plane. The proposed algorithm then assumes that each building has 4 or more facades and that they are mutually perpendicular. Using this assumption, the points are then fitted with a minimum-area bounding rectangle (Freeman and Shapira 1975), defining, in this way, the 4 main façade directions of a building region. The planes of the main façades directions are finally computed considering the same X, Y

coordinates of the bounding rectangle corners and assigning as Z values the mean roof height and the ground mean values, respectively.

4.3.2 Façade extraction from oblique views

The façade regions defined before are used to locate their corresponding image patch on the oblique images, see Figure 29. The images are not relatively oriented by means of photogrammetric procedures but using the raw GNSS/IMU (X,Y,Z,ω,φ,κ) information from the UAV navigation system. The accuracy of such raw GNSS/IMU data can range from 2-10m for the positions and 0.5-5 deg for the attitudes (Eling et al. 2014).

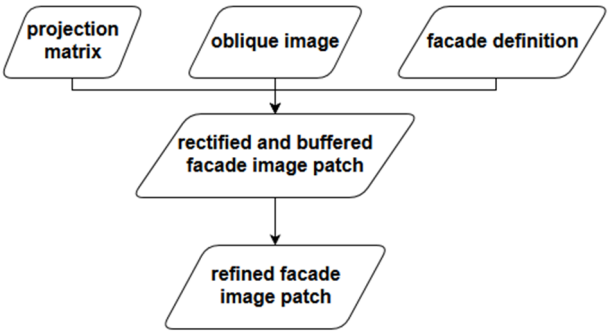

Figure 29 Flowchart regarding the facade extraction from the oblique images

A projection matrix is built using the camera internal parameters and the raw orientation from the UAV stored in the image as *exif* metadata. With the projection matrix and the 4 3D corners of the façade it is possible to re-project the 3D corners into the image. The extracted image patch can then be rectified defining the real-world plane formed by the 4 3D façade corners. However, since the raw UAV image orientation is not accurate, the extraction of the patch containing the whole façade can be a difficult task. The extracted image patch is therefore buffered in image space.

The extracted image patch now contains other objects from the scene in its background, apart from the façade itself. This patch needs to be refined before its use in the damage assessment because: 1) it increases the image area to be analysed; 2) neighbouring objects could also contain damaged areas, hindering the damage classification of the analysed façade. Hence, a further refinement of the façade location is performed using two main sets of information: 1) salient object probability image (Tu et al. 2016), and 2) line segments analysis on the façade (Yi Li and Shapiro 2002).

1) Salient object probability image: the problem to distinguish the façade from its neighbouring objects in image space is in accordance with the objective of salient object detection, which aims to distinguish the *figure* from

the *background* in a given image (Borji et al. 2015). A real-time salient object detection algorithm (Tu et al. 2016), using a minimum spanning tree image representation, is used as one set of information to distinguish the façade from the background resulting from the applied buffer. This salient object detection approach uses the image boundary pixels as seed points for the *background* detection. In this approach, the boundaries of the buffered image patch extracted before are assumed to be dissimilar from the façade. The result of the application of this algorithm is an image containing the probability of a given pixel to belong to the *figure*, in this case, the façade. This probability map is then transformed to a binary image, where only the blob occupying the largest image area is considered.

2) Façade line segments analysis: the images should provide a clear indication of horizontal and vertical elements on the image façade. These lines should appear as perpendicular in the rectified patches. The vertical and horizontal line segments are extracted using the line segment extraction as described in (Košecká and Zhang 2002), which uses the Canny edge detector (Canny 1986) followed by a line fitting stage (Kahn, Kitchen, and Riseman 1990). Line segments which are not vertical nor horizontal (within a 10 degree tolerance) are not considered. In the case the intersection between a vertical and a horizontal line segment is on, or close to the edges of the extended line segments, these are considered as façade line segments (Yi Li and Shapiro 2002).

The salient object detection blob and the façade line segments analysis are finally merged to detect the actual façade within the buffered façade image patch. Every façade line segment which overlays with the salient object blob is considered as part of the façade. The façade area is defined by the image coordinates of the detected façade lines: the maximum and minimum of both *x* and *y* pixel coordinates are used to define the rectangle to crop the façade image patch.

4.3.3 Damage assessment on the refined façade image patch

The cropped façade region is used as input for the damage assessment step. This patch is further divided into equilateral patches (50px size), these are the unit of analysis.

The developed method exploits the presence of vertical and horizontal elements on the rectified patch to quickly analyse the façade. The gradient information has previously been used in contributions aiming at façade decomposition (Recky and Leberl 2010; Teeravech et al. 2014). In this case, the objective is to early select patches in which the gradient information indicates the patches that are candidates for damage. The vertical and horizontal gradients are computed for each patch and posteriorly projected into the horizontal and vertical axes. For each axis, the local maxima and minima

of the gradients are computed, and its concentration per axis is determined (peaks per pixel ratio). Figure 5 contains two examples (one damaged and one non-damaged) of the projection of the vertical and horizontal gradients. The peaks ratio for the non-damaged patch (Figure 5, left) is of 0.11 and 0.12, respectively for the horizontal and vertical projection of the gradients. The peaks ratio for the damaged patch (Figure 5, right), is of 0.45 and 0.33, respectively for the horizontal and vertical projection of the gradients. A candidate for damage is considered when the ratio peaks/pixel is greater than 0.25 on both axes: this number has been experimentally defined and it is intentionally low in order to avoid discarding damaged patches. The image patches where anomalies are detected are further analysed using a pre-trained damage classification CNN as described in Vetrivel et al. (2017). The used model allows to distinguish between damaged and intact regions and it is pre-trained with a set of approximately 8000 training samples (4000 for each class) obtained from several sets of UAV multi-view imagery.

4.4 *Results*

The described method has been applied to the set of data presented in section 2. For each sub-section of the method, a corresponding sub-section in this results section is given.

Figure 30 Projection of the vertical and horizontal gradients :in a non-damaged façade patch (left) and damaged façade patch (right).

4.4.1 Building hypothesis generation and façade definition

This sub-section presents the results for the building detection and façade definition from the sparse point cloud.

Figure 31 presents the sparse point cloud and the corresponding detected buildings (coloured). As can be noted in this figure, the sparse point cloud density is not homogenous throughout the project area, as it highly depends on the texture of the different regions and the image block overlap.

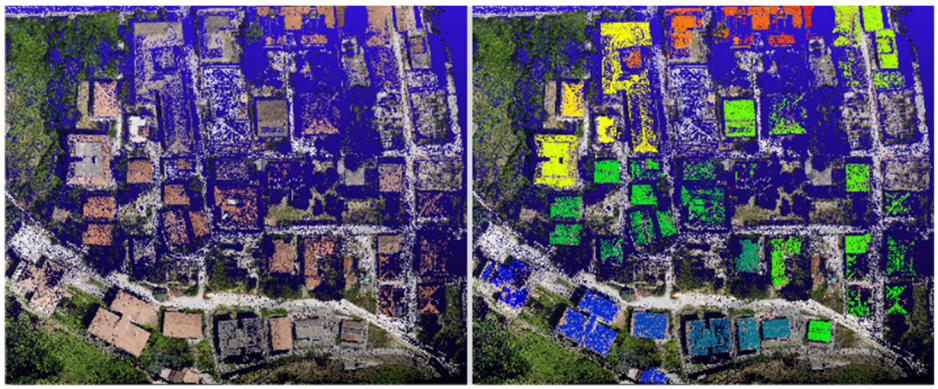

Figure 31 Sparse point cloud, left ; building hypothesis (coloured) overlaid on the sparse point cloud , right

Three examples of the façade definition are given in Figure 32. As can be noted, the proposed approach successfully defines the 4 main façade directions. Since the building edges are usually good candidates for tie points, most of the extracted building regions had a greater concentration of points in those regions. As such, even in the case the point density is low, the main façade identification was successful. This is central to correctly define the minimum bounding rectangle.

With this approach only a building was not identified, because it was partially covered by vegetation, this biased the plane based segmentation and the following building detection. Another issue was the inclusion of points outside the building roof see Figure 33, that happened in one building, hindering the following façade definition.

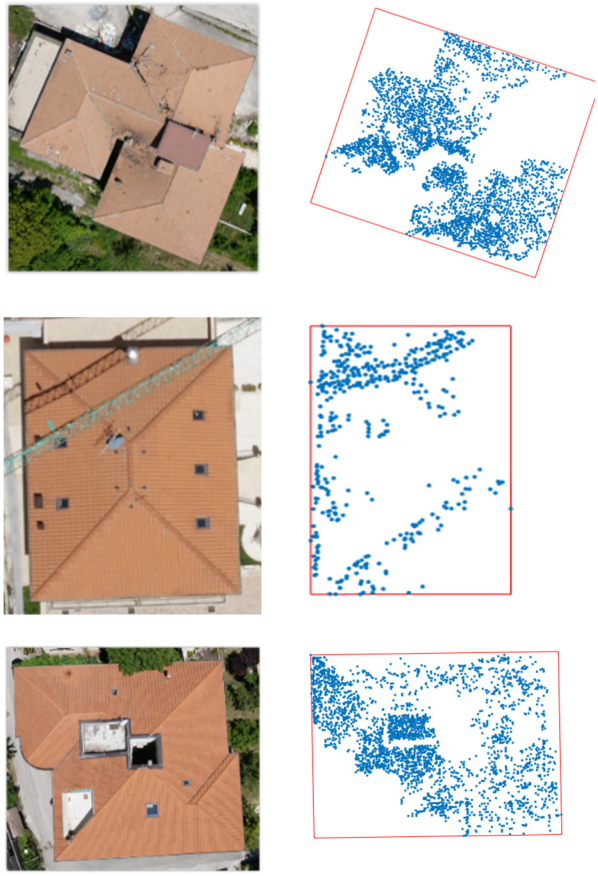

Figure 32 Façade definition. Nadir view of 3 buildings, left and corresponding xy projected sparse points (blue points), and minimum area bounding rectangle (red rectangle), right.

4.4.2 Façade extraction from oblique views

This subsection presents the result of the façade extraction from the oblique images, using the façades defined previously. The used buffer was 350px, to account for the use of the raw orientation coming from the UAV. This buffer was sufficient to successfully capture the whole extent of the façades.

From the 40 considered buffered façade image patches, only 2 were incorrectly extracted due to an incorrect result in the salient object detection (see Figure 34, a and d). This resulted in the extraction of only a small patch of the whole façade. The edges of the buffered image patch in Figure 34 a, contain radiometric similarities with the façade itself. This hindered the real-time salient object detection (since this approach assumes that the image edges are *background* hence a cue to distinguish it from the façade). The façade line segments, in this case, enclosed only a part of the façade.

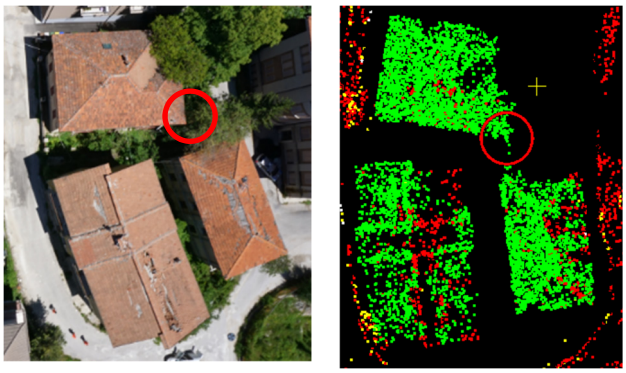

Figure 33 Details of 3 detected building roofs. Left nadir image; right sparse point cloud overlaid with the detected buildings - red circle indicates a segment which is part of the vegetation but is identified as part of a roof segment.

Figure 35 and Figure 36, show the result of the application of the salient object detection combined with the façade line segments to define the façade image patch. In these figures is also visible how the façade line segments information complemented the salient object detection. As it can be noticed in Figure 10, there was no significant impact of the building having more than 4 façades, due to the fitting of the minimum-area bounding rectangle. In this case, and since the other smaller façade shared the same plane orientation, the rectification procedure had a different scale for each façade plane. On the other hand, the results depicted in Figure 36 were hindered by both the presence of façade line segments of a neighbouring façade and by the inclusion of that same façade in the salient object detection. In this case, however, the whole façade patch was still considered.

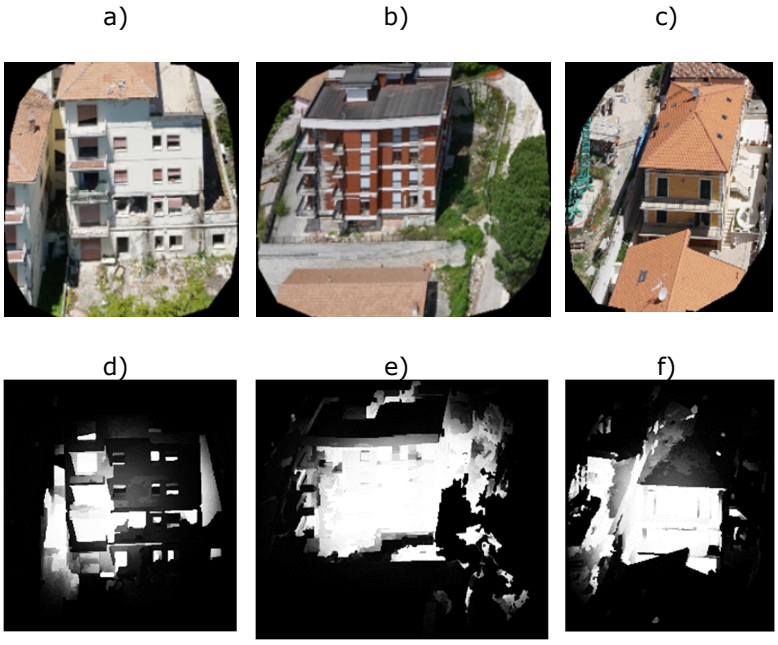

Figure 34 Three examples of the salient object detection results, second row (white regions show a higher probability of the pixel pertaining to the façade)

4.4.3 Damage assessment on the refined façade image patch

This sub-section presents the results for the damage detection on the refined façade image patch.

Table 10 provides the damage classification results, considering the building façades as unit of analysis. Considering 11 damaged façades, 10 contained at least one patch classified as damaged. However, 1 façade was incorrectly classified as not-damaged. Considering the non-damaged façades, 23 were correctly identified as not-damaged, while 6 were incorrectly classified as damaged.

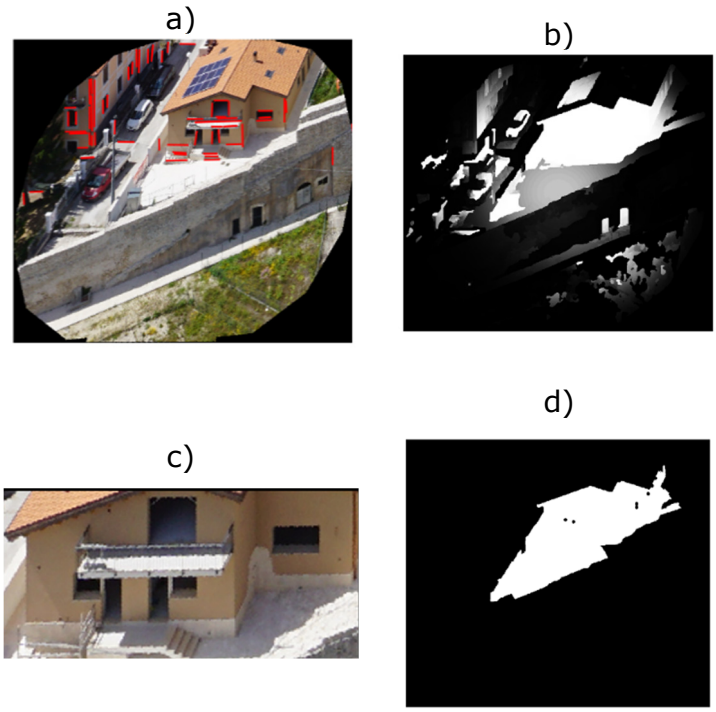

Figure 35 Results of the façade line segments and salient object map: a) façade line segments overlaid in buffered façade patch, b) real-time salient object, c) final refined facade patch, d) binary image of the salient object detection in b)

Table 10 Results of the façade damage classification on 40 façades

Façade damage classification	**No.**
Correctly classified as damaged	10
Incorrectly classified as damaged	6
Correctly classified as not-damaged	23
Incorrectly classified as not-damaged	1
Precision = 62% ; Recall= 90% ; Accuracy= 83%	

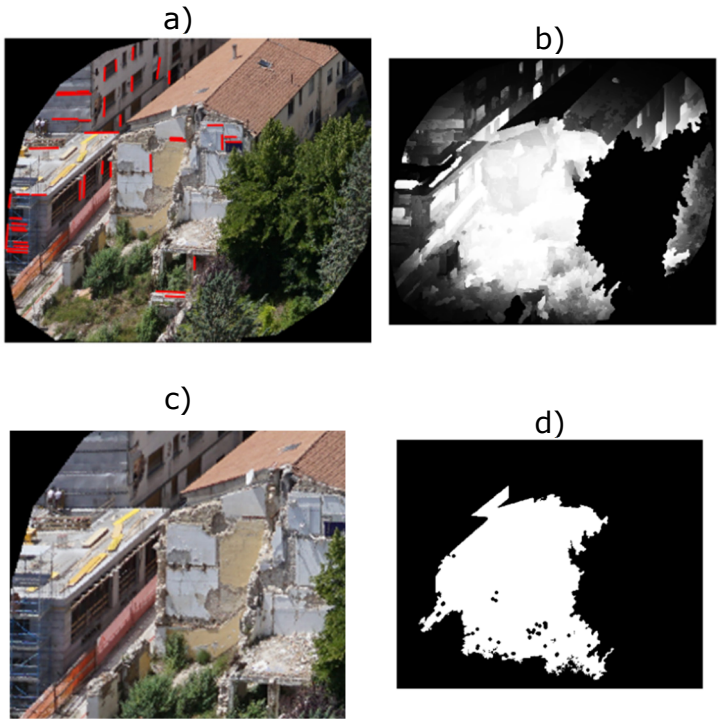

Figure 36 Results of the façade line segments and salient object map: a) façade line segments overlaid in buffered façade patch, b) real-time salient object, c) final refined facade patch, d) binary image of the salient object detection in b)

The visual outcome of the damage classification is depicted in several examples in Figure 37. Ground regions and overhanging elements of the façade contain most of the false positives.

Table 11 provides the results regarding the number of the classified patches on all the façades. The projection of the gradient information in the form of a peaks/pixel ratio allowed to successfully omit 1233 patches from the CNN damage classification. A total of 179 image patches were classified by the CNN, 83 of which as damaged.

Table 11 Results regarding the early selection of patches to be fed to the CNN, considering the 40 façades

Patches assessed for damage	**Patches confirmed damaged (CNN)**	**Patches not considered (gradient peaks)**
179	83	1233

4.5 *Discussion*

The use of the sparse point cloud to extract the buildings, through a plane base segmentation followed by a connected component analysis, has been validated on 40 façades. In spite of the heterogeneous 3D point density in such a point cloud, only one building was not identified due to vegetation occlusions that hindered the plane-based segmentation. However, in cases where the building roof does not reflect the actual orientation of the façades, these are not properly rectified, hindering the consequent analysis.

The buffer used in the extraction of the façade image patch also sufficed to account for the poor raw orientation from the UAV navigation system. However the adoption of the same buffer size for every façade is not optimal due to the variability in the image georeferencing inaccuracies and due to the varying façade size.

The posterior façade patch refinement using line segments and the salient object image, successfully depicted the façade location. However, 2 façades were incorrectly extracted due to a wrong salient object detection.

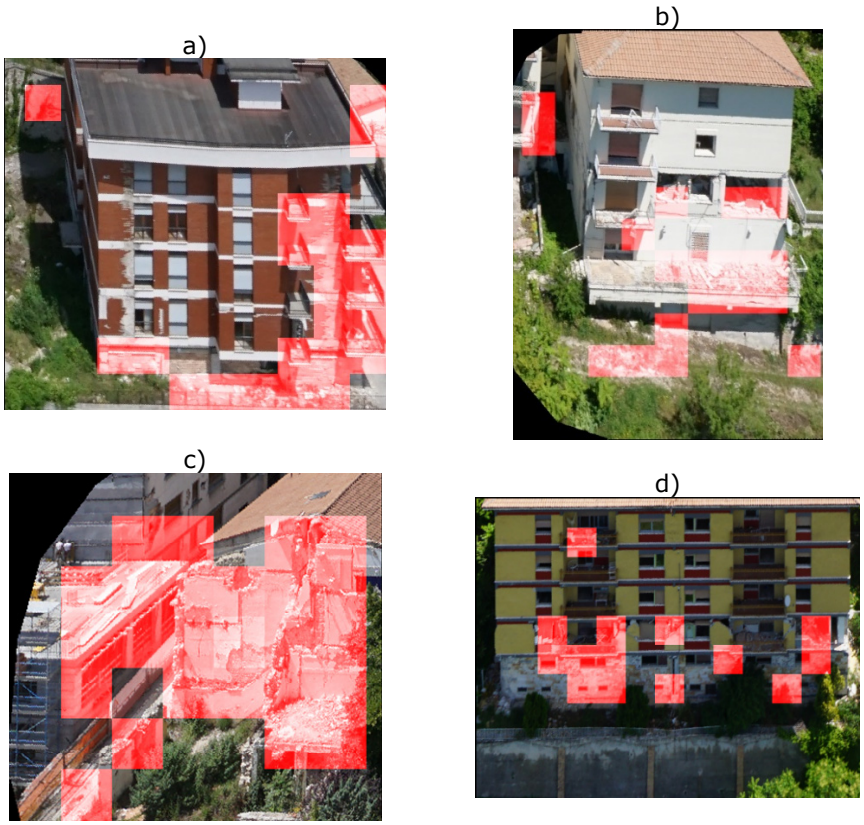

Figure 37 Refined façade damage detection results: a, b, c and d. Damaged patches overlaid in red.

The use of the projection of the vertical and horizontal gradient, allowed to decrease the refined façade image patch regions to be processed by the damage classification step. Only approximately 1/6 of the total regions contained in the refined façade image patch were considered for classification.

The results of the damage classification using a CNN, at a refined façade image patch level, are in accordance with the results obtained in Vetrivel et al. (2017). Scene characteristics like ground regions, overhanging objects in the façade, construction sites and roof tiles, are the main cause of the false positives (6) reported in the results.

In spite of the increase in efficiency it must be noted that the façades which were wrongly defined from the sparse point cloud or incorrectly extracted from the images, are not assessed for damage. This is one of the main drawbacks of the proposed method.

4.6 *Conclusions*

In this chapter a methodology to increase the efficiency of façade damage detection using a set of UAV multi-view imagery was presented. The higher productivity of the method was achieved by reducing the number of images and image regions to be analysed for damage in a three step approach.

One of the major contributions of the presented approach was the possibility of using the sparse point cloud to detect building roofs. This allowed to omit the generation of the computationally expensive DIM, increasing the speed of the façade damage detection.

The 4 main façade directions, together with the raw orientation information from the navigation system of the UAV, were used to identify the façades in the oblique images. Due to the uncertainties of such orientation information, a wide image buffer was adopted. Future work will address this issue, by relating, the façade size with the size of the buffer to apply.

The salient object detection coupled with the façade line segments, successfully identified the façade in the buffered image patch, reducing the area to be used in the subsequent damage classification step.

The damage detection using the CNN approach gave 6 false positives. The performances of the CNN for this step will be addressed in a future work by re-designing the network (as suggested in Cheng et al. (2017)) and by extending the used training dataset. In this regard, the reduced number of post-earthquake UAV multi-view datasets could represent a limiting factor. Another possibility to improve these results would be to consider more than one image to assess the damage state of a given façade.

The presented methodology is still an on-going work, the final goal would be to reach a near-real-time façade damage detection. In this regard, a new way

to acquire images could be considered, planning the acquisitions of the oblique views on the basis of the buildings extracted from the sparse point cloud, hence decreasing the amount of collected images. Moreover, the information provided by nadir images may be also used to detect evidences of façade damage, such as blow out debris or rubble piles in the vicinity of the building. This would enable a prioritization of the planed oblique views.

4.7 *References of Chapter 4*

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

Armesto-González, J., Riveiro-Rodríguez, B., González-Aguilera, D., & Rivas-Brea, M. T. (2010). Terrestrial laser scanning intensity data applied to damage detection for historical buildings. *Journal of Archaeological Science*, *37*(12), 3037–3047. https://doi.org/10.1016/j.jas.2010.06.031

Axelsson, P. (2000). DEM generation from laser scanning data using adaptive TIN models. *International Archives of Photogrammetry and Remote Sensing*, *33*, 111–118. Amsterdam.

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2015). Salient object detection: a benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5706–5722. https://doi.org/10.1109/TIP.2015.2487833

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. *Proceedings of the IEEE*, 1–19. https://doi.org/10.1109/JPROC.2017.2675998

Dell'Acqua, F., & Gamba, P. (2012). Remote sensing and earthquake damage assessment: experiences, limits, and perspectives. *Proceedings of the IEEE*, *100*(10), 2876–2890. https://doi.org/10.1109/JPROC.2012.2196404

Dell'Acqua, F., & Polli, D. A. (2011). Post-event only VHR radar satellite data for automated damage assessment. *Photogrammetric Engineering & Remote Sensing*, *77*(10), 1037–1043. https://doi.org/10.14358/PERS.77.10.1037

Dong, L., & Shan, J. (2013). A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, *84*, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011

Eling, C., Klingbeil, L., Kuhlmann, H., Bendig, J., & Bareth, G. (2014). *A precise direct georeferencing system for UAVs*. https://doi.org/10.5880/TR32DB.KGA94.6

Fernandez Galarreta, J., Kerle, N., & Gerke, M. (2015). UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. *Natural Hazards and Earth System Science*, *15*(6), 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

Freeman, H., & Shapira, R. (1975). Determining the minimum-area encasing rectangle for an arbitrary closed curve. *Communications of the ACM*, *18*(7), 409–413. https://doi.org/10.1145/360881.360919

Gerke, M., & Kerle, N. (2011). Automatic structural seismic damage assessment with airborne oblique Pictometry© imagery. *Photogrammetric Engineering & Remote Sensing*, *77*(9), 885–898. https://doi.org/10.14358/PERS.77.9.885

Gokon, H., Post, J., Stein, E., Martinis, S., Twele, A., Muck, M., … Matsuoka, M. (2015). A method for detecting buildings destroyed by the 2011 Tohoku earthquake and tsunami using multitemporal TerraSAR-X data. *IEEE Geoscience and Remote Sensing Letters*, *12*(6), 1277–1281. https://doi.org/10.1109/LGRS.2015.2392792

Kahn, P., Kitchen, L., & Riseman, E. M. (1990). A fast line finder for vision-guided robot navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(11), 1098–1102. https://doi.org/10.1109/34.61710

Khoshelham, K., Oude Elberink, S., & Sudan Xu. (2013). Segment-Based classification of damaged building roofs in aerial laser scanning data. *IEEE Geoscience and Remote Sensing Letters*, *10*(5), 1258–1262. https://doi.org/10.1109/LGRS.2013.2257676

Košecká, J., & Zhang, W. (2002). Video Compass. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *Computer Vision — ECCV 2002* (Vol. 2353, pp. 476–490). https://doi.org/10.1007/3-540-47979-1_32

Ma, J., & Qin, S. (2012, July). *Automatic depicting algorithm of earthquake collapsed buildings with airborne high resolution image*. 939–942. https://doi.org/10.1109/IGARSS.2012.6351400

Marin, C., Bovolo, F., & Bruzzone, L. (2015). Building Change Detection in Multitemporal Very High Resolution SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(5), 2664–2682. https://doi.org/10.1109/TGRS.2014.2363548

Recky, M., & Leberl, F. (2010, July). *Window detection in complex facades*. 220–225. https://doi.org/10.1109/EUVIP.2010.5699128

Sui, H., Tu, J., Song, Z., Chen, G., & Li, Q. (2014). A novel 3D building damage detection method using multiple overlapping UAV images. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XL–7*, 173–179. https://doi.org/10.5194/isprsarchives-XL-7-173-2014

Teeravech, K., Nagai, M., Honda, K., & Dailey, M. (2014). Discovering repetitive patterns in facade images using a RANSAC-style algorithm.

ISPRS Journal of Photogrammetry and Remote Sensing, *92*, 38–53. https://doi.org/10.1016/j.isprsjprs.2014.02.018

Tu, W.-C., He, S., Yang, Q., & Chien, S.-Y. (2016, June). *Real-time salient object detection with a minimum spanning tree*. 2334–2342. https://doi.org/10.1109/CVPR.2016.256

United Nations. (2015). *INSARAG Guidelines, Volume II: Preparedness and Response, Manual B: Operations*. United Nations Office for the Coordination of Humanitarian Affairs (OCHA).

Vetrivel, A., Duarte, D., Nex, F., Gerke, M., Kerle, N., & Vosselman, G. (2016). Potential of multi-temporal oblique airborne imagery for structural damage assessment. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *III–3*, 355–362. https://doi.org/10.5194/isprsannals-III-3-355-2016

Vetrivel, A., Gerke, M., Kerle, N., & Vosselman, G. (2016). Identification of structurally damaged areas in airborne oblique images using a Visual-Bag-of-Words approach. *Remote Sensing*, *8*(3), 231. https://doi.org/10.3390/rs8030231

Vetrivel, Anand, Gerke, M., Kerle, N., Nex, F., & Vosselman, G. (2017). Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS Journal of Photogrammetry and Remote Sensing*. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Vosselman, G. (2012). Automated planimetric quality control in high accuracy airborne laser scanning surveys. *ISPRS Journal of Photogrammetry and Remote Sensing*, *74*, 90–100. https://doi.org/10.1016/j.isprsjprs.2012.09.002

Yi Li, & Shapiro, L. G. (2002). *Consistent line clusters for building recognition in CBIR*. *3*, 952–956. https://doi.org/10.1109/ICPR.2002.1048195

5 Potential of multi-temporal oblique airborne imagery for structural damage assessment[4]

[4] This chapter is based on the article:
Vetrivel, A., D. Duarte, F. Nex, M. Gerke, N. Kerle, and G. Vosselman. 2016. "Potential of multi-temporal oblique airborne imagery for structural damage assessment." ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III–3 (June): 355–62. https://doi.org/10.5194/isprsannals-III-3-355-2016.

5.1 *Introduction*

Damage assessment is an imperative process to be carried out immediately after the disaster event for effective planning and execution of response and recovery actions. Assessing building damages over large areas affected by hazard events with ground observations is not efficient. Alternatively, remote sensing-based approaches have been recognized as useful means for assessing synoptic building damage. Detailed information of an affected area can be provided in a short time using a variety of sensors such as optical, SAR and LiDAR (Khoshelham et al., 2013; Maruyama et al., 2014; Miura et al., 2007). In particular airborne oblique images have been recognized as a valuable data source to assess building damages because, compared to traditional nadir views, they allow the complete inspection of the external outlines of the building, such as roofs and façades (Murtiyoso et al., 2014). Nowadays, airborne images are captured with high overlap, and the generated point clouds can be exploited in the damage assessment process as well (Sui et al., 2014). Geometrical deformations such as partial/complete collapse, pancake collapse, inclination, broken and dislocation of elements can be easily derived by 3D geometric information (Fernandez Galarreta et al., 2015), while damages such as cracks and spalling can be inferred from the images directly. Several papers have highlighted the potential of synergistic use of 3D point cloud and images for building damage assessment (Gerke and Kerle, 2011; Vetrivel et al., 2015). However, only few studies have looked at the use of digital oblique aerial imagery for structural damage assessment, and were focused on (mono-temporal) post-event information (Gerke and Kerle, 2011; Vetrivel et al., 2015). The major limitation of this approach is that damage is inferred based on a set of ontological assumptions: i.e. a surface with unusual radiometric or geometric characteristics is assumed to be damaged, while manmade objects are assumed to have a regular shape and uniform radiometric characteristics. These assumptions have limitation in complex environments, leading to a high rate of false alarms, which reduces their reliability and operational utility. In Vetrivel et al. (2015), damages presenting regular and uniform shapes (false negative), or intact regions characterized by cluttered and non-uniform radiometric distributions (false positive), were incorrectly classified due to these assumptions. The above uncertainties can be alleviated if pre-event data are available for reference. Many studies have demonstrated the potential of multi-temporal data for damage assessment, though with most focusing on nadir-view images (Dong and Shan, 2013; Murtiyoso et al., 2014). To our knowledge, no methods have been reported yet for identifying building damages, namely façades, using multitemporal oblique images and/or 3D point clouds. In this chapter the first implementation of an automated algorithm, for building damage assessment focusing on the façades, from multi-temporal oblique images is presented. Although geometrically more stable cameras are used nowadays in oblique airborne systems, many data sets are captured with

less sophisticated camera systems, and image overlap is often restricted to 2-fold. Hence, for such configurations one has to cope with dense image matching point clouds of minor quality (relatively large random error margin, gaps). In particular façade regions are generally represented by sparse and very noisy 3D points, as they are more cluttered and often occluded (Rupnik et al., 2014). The proposed method uses the point cloud to locate the façades and focus on image information to assess the damage of a given façade.

The chapter addresses the analysis of multi-temporal oblique images for identifying the damages along façades which are often not well modelled in the generated point clouds.

The detailed description of the methodologies and the results achieved on the test area of L'Aquila (Italy) will be presented in detail.

5.2 *Data description*

The data used are corresponding to the city of L'Aquila, Italy in which an earthquake occurred on 6th April 2009. The data consist of two airborne oblique acquisitions (August 2008 and May 2009) covering the city with both oblique (4 cameras) and nadir (1 camera) imagery, captured by small format DSLR cameras. Images were acquired at a flying height of approximately 1000 m allowing for an average ground sampling distance of 16 cm on oblique views. The flight was conducted considering a forward overlap between 60-70% and side overlap between 35-45%, allowing to derive a 3D point cloud. The registration was achieved computing tie-points from all the imagery, forcing both epochs to share a local coordinate system. Dense image matching was then performed separately on both epochs.

5.3 *Method*

The objective of the following method is to automatically detect building façade changes by comparing their radiometric values. It will lay the ground for further developments focusing first on the extraction and rectification of the image patches containing the façades, followed by the comparison itself, and considering three main categories: highly damaged or collapsed façades, lower levels of damage (changes), and undamaged buildings.

To perform the multi-temporal comparison between the façades, these building elements must be, beforehand, extracted from the images. The pre-event 3D point cloud allows the identification and extraction of the points relative to the façades. These can be back-projected into the image, using the correspondent projection matrices, defining the boundaries of the image patches to extract. The 3D points corresponding to the façades will also be used to define the plane containing the façade by fitting a least square plane. Using this 3D plane and the extracted façade patches, these can be rectified using a homography

matrix. An interpolation is performed on the gaps produced by the projection of the pixels to the real world façade plane (see Figure 12). Variable resolution and brightness can be detected according to the point of view of each image. These problems affect the results independently of the epoch of the images (same or different epoch). The comparison itself is made by determining the correlation coefficient between the rectified façade patches. This correlation coefficient is computed using a 7 by 7 pixels moving window, determining local (on each window position) and global (mean of the considered façade) values of the computed cross correlation. Nevertheless, only results of the inter-epoch correlation are not sufficient, since they do not have an actual meaning of change/no change, but just provide a correlation value of the pixels of the compared façades. To normalize the correlation values and increase the feasibility in the change detection, the correlation coefficient is first performed using different images of the same epoch as this value serves as reference to judge the multi-epoch comparison. The façades with a difference between intra- and inter-epoch correlation coefficients bigger than an imposed value will be considered as highly damaged or collapsed buildings. Again, this imposed value to limit the difference in the correlation values between epoch is based on the intra-epoch correlation results (see 4.2 Results). The same for undamaged buildings where similar correlation coefficients in both the intra- and inter-epoch indicate the presence of an undamaged element. The correlation values from the intermediate category, lower levels of damage, are still the most critical to be automatically interpreted, and they are classified as changes in the current method implementation.

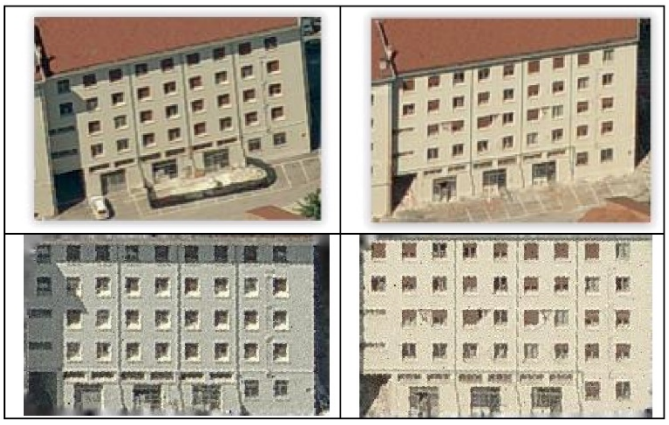

Figure 12. Example of two pre- and post-event subsets of oblique images containing the façade (above) and respective rectified images (below).

5.4 *Results*

This section presents the results according to the categories defined earlier. As explained before the intra-epoch radiometric comparison will serve as

reference value for the inter-epoch comparison. The undamaged case will be assessed first (Figure 38). It will be followed by an example posing the possible intraepoch differences between the extracted image patches from two distinct images (Figure 39), given the problems addressed in the previous section. The inter-epoch comparison will then be addressed considering damage related changes (Figure 40) and other changes (Figure 41). Finally, the collapsed building case will be depicted (Figure 42). Considering the façade presented in Figure 13, the correlation coefficient was 0.78 intra-epoch and 0.75 inter-epoch. This similarity can categorize this façade as unchanged and consequently not damaged.

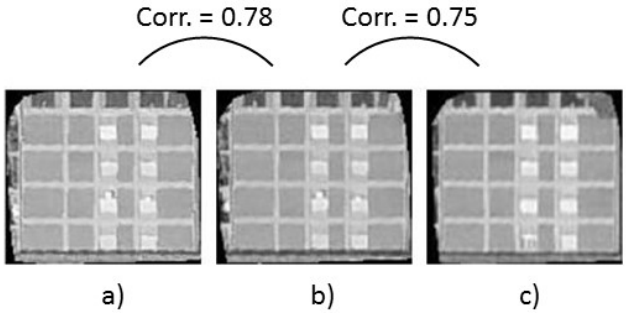

Figure 38 Same façade extracted from both epochs. a) and b) relative to pre-event and c) post event.

Figure 39 represents another façade element in which the intraepoch correlation coefficient is lower than in the former example. The balconies which are not in the defined plane, are hence consequently wrongly rectified. Different illumination settings are also noticeable on the shadows of the shown rectified patches. The global correlation value on the façade is therefore very low (0.51) and the correlation values are extremely low (darker areas) in correspondent balconies and associated shadows (Figure 39, rightmost part)

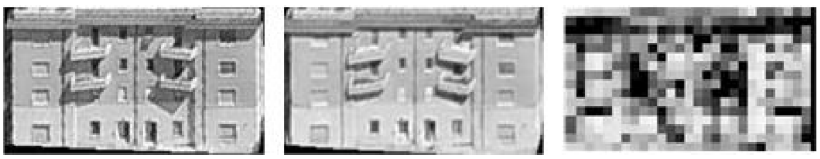

Figure 39 Pre-event rectified image patches and corresponding correlation coefficient.

In Figure 40, the intra-epoch correlation coefficient is 0.52 and inter-epoch correlation is 0.40. In this case low correlation values are mainly due to the lack of texture in a large portion of the façade and different position of the shadows in the images; in the inter-epoch case, the value is further decreased by the presence of a large spalling area on the façade.

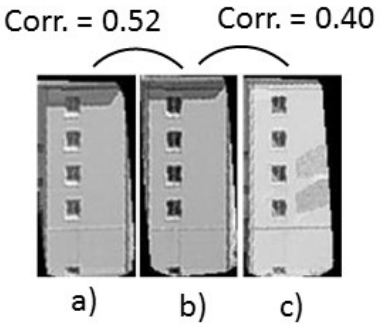

Figure 40 Hazard-related changes. Same façade extracted from both epochs. a) and b) relative to pre-event and c) post event.

Not all changes are damage related as can be seen in Figure 41. Here the intra- and inter-epoch correlation coefficients were 0.66 and 0.33, respectively. Unlike the latter case, the changes which decreased the correlation coefficient are not hazard related and are due to the removal of banners present in the pre-event.

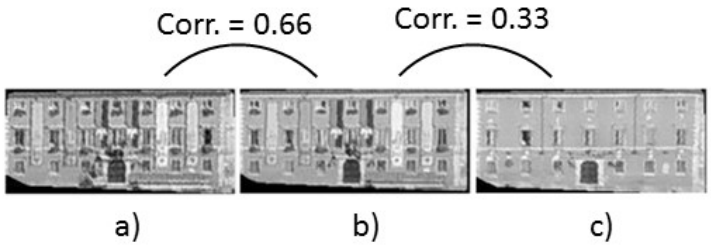

Figure 41 Changes not hazard related. Same façade extracted from both epochs. a) and b) relative to pre-event and c) post event.

However, correlation values are completely different in the case of complete collapses. The façade shown Figure 39 is considered in an inter-epoch comparison: the mean of the correlation coefficient drops drastically (from 0.51 to 0.04) indicating directly the presence of a collapsed building (Figure 42). This can be performed for several façades to confirm the outcome of the categorization, as collapsed, for the whole building.

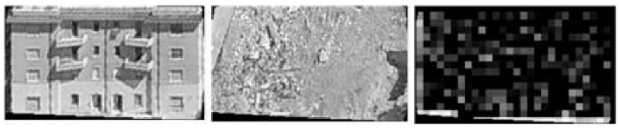

Figure 42 Total collapse example, rectified images on both epochs and correlation coefficient matrix.

5.5 *Discussion*

The results obtained above allow to confirm that the present methodology can differentiate between the three proposed categories, collapsed/highly damage, presence of lower levels of damage and undamaged buildings, using a computationally light approach. In the collapsed or highly damaged case this can even be confirmed using the available façade elements and also roof elements corresponding to a same building. Considering the undamaged/unmodified case, the correlation coefficient similarity will mostly indicate the presence of the same unchanged element. Although, as seen in Figure 14. Pre-event rectified image patches and corresponding correlation coefficient., the presence of balconies or other overhanging details will decrease the correlation between facades, since these were assumed flat in order to perform the image rectification. Analogously a very low correlation value can immediately indicate a high level of damage. Concerning the intermediate category, where the changes happened at a façade level, the definition of a correlation interval which includes these elements may not be so direct like the previously referred categories; nonetheless, the multi-temporal component can certainly aid in the definition of such interval.

5.6 *Conclusion and outlook*

The presented methodology aimed at comparing building façades at an image level, in order to infer the presence of damages on them. A rough distinction between the three proposed damaged levels using a fast approach was demonstrated. However, due to the variability of light conditions and different point of views, the correct selection of damages on the façades still remains a challenge. The façades that are present in just an epoch will have to be carefully assessed. These variations in the occlusion can have its origin in the data acquisition itself, another example can be vegetation changes or the modification of the urban configuration. A segmentation of the building façade (in 2D and 3D) will be performed on the façades in order to detect and remove windows and balconies, restricting the change search on the facade walls. The shadow detection will be addressed as well. This method does not need the computation of point clouds from different epochs but only co-registered images. Already existing 3D city models could be used to define (and rectify) the façade position, strongly reducing the time needed in the damage map generation since there is no need to generate a point cloud. It would also allow not only the integration of the damage results with the city model itself but also to ease an integration with damage maps from other sources.

5.7 *References of Chapter 5*

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS Journal of Photogrammetry and Remote Sensing 84, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011

Fernandez Galarreta, J., Kerle, N., Gerke, M., 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Natural Hazards and Earth System Science 15, 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

Gerke, M., Kerle, N., 2011. Automatic structural seismic damage assessment with airborne oblique Pictometry© imagery. Photogrammetric Engineering & Remote Sensing 77, 885–898. https://doi.org/10.14358/PERS.77.9.885

Khoshelham, K., Oude Elberink, S., Sudan Xu, 2013. Segment-based classification of damaged building roofs in aerial laser scanning data. IEEE Geoscience and Remote Sensing Letters 10, 1258–1262. https://doi.org/10.1109/LGRS.2013.2257676

Maruyama, Y., Tashiro, A., Yamazaki, F., 2014. Detection of collapsed buildings due to earthquakes using a digital surface model constructed from aerial images. Journal of Earthquake and Tsunami 08, 1450003. https://doi.org/10.1142/S1793431114500031

Miura, H., Yamazaki, F., Matsuoka, M., 2007. Identification of damaged areas due to the 2006 central Java, Indonesia earthquake using satellite optical images. IEEE, pp. 1–5. https://doi.org/10.1109/URS.2007.371867

Murtiyoso, A., Remondino, F., Rupnik, E., Nex, F., Grussenmeyer, P., 2014. Oblique aerial photography tool for building inspection and damage assessment, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 309–313. https://doi.org/10.5194/isprsarchives-XL-1-309-2014

Rupnik, E., Nex, F., Remondino, F., 2014. Oblique multi-camera systems orientation and dense matching issues. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-3/W1, 107–114. https://doi.org/10.5194/isprsarchives-XL-3-W1-107-2014

Sui, H., Tu, J., Song, Z., Chen, G., Li, Q., 2014. A novel 3D building damage detection method using multiple overlapping UAV images. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL–7, 173–179. https://doi.org/10.5194/isprsarchives-XL-7-173-2014

Vetrivel, A., Markus Gerke, Norman Kerle, George Vosselman, 2015. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images, in: ISPRS Journal of

Photogrammetry and Remote Sensing. pp. 61–78. https://doi.org/10.1016/j.isprsjprs.2015.03.016

6 Detection of seismic façade damages with multi-temporal aerial oblique imagery[5]

[5] This chapter is based on the article:

Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Damage detection on building façades using multi-temporal aerial oblique imagery. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., 2019, IV-2/W5, 29-36

Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Detection of seismic façade damages with multi-temporal oblique aerial imagery (submitted for review)

Abstract

Remote sensing images have long been recognized as useful for the detection of building damages, mainly due to their wide coverage, revisit capabilities and high spatial resolution of the used sensors. The majority of contributions aims at identifying debris and rubble piles, as the main focus is to assess collapsed and partially collapsed structures. However, these approaches might not be optimal for the image classification of façade damages, where damages might appear in the form of spalling, cracks and collapse of small segments of the façade. Only a few studies focus their damage detection on the façades using only post-event images. A multi-temporal approach is missing. One of the main objectives of the chapter is to optimally merge pre- and post-event aerial oblique imagery within a supervised classification approach using convolutional neural networks to detect façade damages. The second objective is related to the fact that façades are normally depicted in several views in aerial manned photogrammetric surveys; hence, different procedures combining these multi-view image data are also proposed and embedded in the image classification approach. Six multi-temporal approaches are compared against 3 mono-temporal ones. The results indicate the superiority of multi-temporal approaches (up to ~25% in f1-score) when compared to the mono-temporal ones. The best performing multi-temporal approach takes as input sextuples (3 views per epoch, per façade) within a late fusion approach to perform the image classification of façade damages. However, the detection of smaller evidences of damage, such as smaller cracks or smaller areas of spalling, remains challenging in this approach, mainly due to the low resolution (~0.14m ground sampling distance) of the used dataset.

6.1 Introduction

Earthquakes are the deadliest natural hazard, and are responsible for almost a quarter of the recorded economic losses by disasters in the last 20 years (Wallemacq and House, 2018). The built-up environment plays a major role in both of the latter issues, where a growing migration to megacities is further increasing the risk associated with earthquakes (Dong and Shan, 2013). A synoptic assessment of the damaged buildings over an affected region is therefore useful in the several steps of the disaster management cycle. The localization of collapsed and partially collapsed buildings is mandatory for an efficient deployment of first responders immediately after an event occurs (United Nations, 2015). On the other hand, the thorough damage assessment of a building can be also valuable for recovery and insurance purposes (United Nations, 2009) performed at a later stage of the disaster management cycle.

The manual inspection of damaged buildings is a time and resource consuming procedure, aggravated by the post-disaster context. Many approaches using remote sensing have been proposed for building damage assessment at several

scales and with different platforms and sensors. Satellite, aerial and terrestrial platforms coupled with optical (Curtis and Fagan, 2013; Cusicanqui et al., 2018; Dubois and Lepage, 2014; Sui et al., 2014), radar (Brunner et al., 2011; Jung et al., 2018; Li et al., 2012) or laser instruments (Armesto-González et al., 2010; Khoshelham et al., 2013) have already been proposed as main source of data for building damage assessment. However, the largest effort has been focused on the methods using optical images as input (Cusicanqui et al., 2018; Dell'Acqua and Gamba, 2012; Duarte et al., 2018a; Dubois and Lepage, 2014; Vetrivel et al., 2017). This is due to several factors, among them the availability of images being collected by satellite and aerial platforms, when compared with laser scanners for example, and their frequent use in photogrammetric processes to generate 3D models (Gerke and Kerle, 2011; Vetrivel et al., 2017).

Many approaches have been proposed to detect damaged regions in remote sensing imagery (Duarte et al., 2018a; Fernandez Galarreta et al., 2015; Gerke and Kerle, 2011; Sui et al., 2014; Vetrivel et al., 2017). Often these approaches rely on features extracted from images which are later used as input for a given classifier. Convolutional neural networks (CNN) have been shown to outperform the image classification with traditional handcrafted features in many applications (Krizhevsky et al., 2012; Long et al., 2015), and this has been confirmed in the detection of building damages in remote sensing images (Duarte et al., 2018a; Vetrivel et al., 2016), too.

Most of the recent image-based damage detection frameworks rely on CNN to determine if a given image patch contains a damage region in a binary classification approach (Duarte et al., 2018a; Nex et al., 2019; Vetrivel et al., 2017). Such frameworks were designed to detect rubble piles and/or debris from satellite (Duarte et al., 2018b) and aerial images (Vetrivel et al., 2017). The details visible in satellite images and the (near) nadir view only allow a rough analysis and identification of collapsed buildings (Kerle and Hoffman, 2013). In contrast, aerial (manned and unmanned) systems have a higher spatial resolution and may also capture oblique views. Smaller details of damage evidences may be therefore identified, such as spalling and cracks (see Figure 43).

Moreover, aerial images are usually captured with enough overlap to derive a 3D point cloud through dense image matching. Depending on the data resolution 3D models may be the input to detect geometrical deformations of the built environment (Gerke and Kerle, 2011), while the images may be used to detect rubble piles and/or debris (Vetrivel et al., 2017), as well as smaller signs of damage such as cracks (Fernandez Galarreta et al., 2015).

In the literature most of the contributions consider the detection of partially or completely collapsed buildings. Given that these are trained with image samples containing rubble piles and debris, these are not optimal for the façade

case (Duarte et al., 2017). The specific case of façade damage detection is only discussed in a few contributions. Fernandez Galarreta et al. (2015) extracted cracks and spalling from façades from unmanned aerial vehicle (UAV) imagery, relying both on the image and 3D features. Gerke and Kerle (2011) used multi-view aerial imagery and derived a 3D point cloud to extract features and identify damaged buildings, and at the same time classified the damage of a given building into three classes, based on the European Macroseismic Scale (EMS-98). More recently, Tu et al. (2017) identified damaged façades using local symmetry features and the Gini Index extracted from aerial oblique images. The authors assumed symmetric façades and considered the deviations from that symmetry to be façade damage proxies. Furthermore, only two contributions used pre- and post-event multi-view aerial imagery in a multi-temporal approach to detect damaged façades. Vetrivel et al. (2016) tested the potential of multi-temporal aerial imagery by using a correlation coefficient to determine the similarity between two rectified façade image patches. Duarte et al. (2019), reported preliminary results regarding the use of a supervised classifier to detect damaged façades using multi-temporal oblique imagery. The authors used two different approaches to merge the multi-temporal oblique image data, which clearly outperformed mono-temporal approaches. Nonetheless, the results only achieved ~66% accuracy in the best performing multi-temporal approach.

The use of airborne oblique imagery has substantially increased in the last decade, allowing the efficient collection of detailed high-resolution information over urban areas. Aerial surveys are regularly performed in many countries and enable their use to detect changes over time and after sudden events.

Figure 43. Examples of nadir images depicting rubble piles and debris, left. Damaged façades shown in oblique imagery, right.

Exploiting the availability of multi-temporal datasets, six different approaches to detect façade damages from pre- and post-event are discussed. Three mono-temporal approaches (using only post-event data) are used as reference.

The focus on the multi-temporal experiments is twofold:

1. To determine the optimal approach to merge the multi-temporal information within deep learning framework for the image classification of façade damages;
2. To leverage the redundancy present in aerial (manned) surveys to extract several façade image patches per façade in each epoch, and to combine these within the frameworks presented in 1).

An additional effort is made to conceive methods exploiting only image information and pre-event 3D models to be (potentially) used in near-real time conditions (assuming the availability of multi-temporal data), when fast and automated methods are needed.

The following section presents a short background. The datasets used in the experiments are presented in section 3. This section also addresses the façade

extraction from the aerial oblique imagery. Section 4 presents the methodology for the multi-temporal image classification of façade damages. Section 5 presents the experiments and results, which are followed by the discussion and conclusions.

6.2 *Background*

This sub-section focuses on CNN and its role in multi-temporal studies using remote sensing imagery. It starts with a brief description of recent developments in CNN that were adapted to this work. An overview of multi-temporal approaches using remote sensing imagery is also given.

Supervised deep learning methods have become an established machine learning technique for image-based tasks, where CNN play a central role. CNN usually achieve high discriminative capacity by stacking convolutions in a hierarchical manner, learning from lower level features to higher levels of abstraction (Krizhevsky et al., 2012). However, in this way each layer is only connected with the previous and posterior layer. Hence, there is feature information that may be lost during backpropagation, especially from earlier layers (Yu et al., 2017). To tackle this, short connections between non-adjacent layers started to be used (He et al., 2016; Huang et al., 2017). He et al. (2016), introduced the concept of residual connection, in which the authors used short connections through element-wise addition of non-consecutive layers. This allowed for the use of deeper networks while maintaining their efficiency, which is often translated into more accurate predictions.

More recently it was found to be preferable to concatenate the feature information instead of performing element-wise addition. Huang et al. (2017) proposed the densely connected convolutional network, introducing short connections in the form of the concatenation of feature maps. This difference allows the model to be more compact, given that every layer receives feature information from the layers preceding it. Thus, features of a given layer may be re-used in later stages of the network, which offers them more representability.

Another aspect of CNN that is closely related with remote sensing is the use of dilated convolutions. These were proposed by Yu and Koltun (2017) and are convolutions with a kernel with pre-defined gaps. This is translated into a wider receptive field, capturing more contextual information. Given that the receptive field of the dilated convolutions is larger, it can capture features over a larger image region, while maintaining a low number of parameters due to the gapped kernel (Yu et al., 2017). This has been taken advantage of by researchers in remote sensing image recognition tasks, who extensively used dilated convolutions in remote sensing tasks (Hamaguchi et al., 2017; Jiang and Lu, 2018; Persello and Stein, 2017; Zhang et al., 2019).

The remote sensing community has been adapting and proposing CNN approaches for the singularities of earth observation tasks and data. For example, such CNN have been directly used in image classification (F. Hu et al., 2015; W. Hu et al., 2015; Maggiori et al., 2017; Nogueira et al., 2017) and image segmentation (Kampffmeyer et al., 2016; Längkvist et al., 2016; Volpi and Tuia, 2017) approaches. However, CNN have for example also been used to merge different modalities of remote sensing data (e.g., 3D and images) (Audebert et al., 2018, 2017; Duarte et al., 2018a), annotate aerial images (Xia et al., 2015; Zhuo et al., 2019) and perform multi-temporal studies (Daudt et al., 2018; Jung et al., 2018; Wang et al., 2018; Zhang et al., 2019).

Multi-temporal studies using CNN often focus their attention on the optimal merging of the different epochs of imagery. Several approaches have been proposed, mostly using satellite imagery and nadir constrained images. Wang et al. (2018) reported that for the task of change detection in satellite imagery it would be preferable to consider the subtraction of pre- and post-event imagery, with the new image being then fed to the CNN. Daudt et al. (2018) tested two approaches to detect changes in multi-temporal satellite imagery. One of the approaches considered two branches of convolutional layers with shared weights (also known as Siamese network), one for each epoch, while the other considered a single set of convolutions performed on the concatenation of the pre- and post- event data as the first stage of the network. The authors reported that early fusion of the inputs was preferable for the detection of changes from satellite imagery. In a different study with the objective of the detection of landslides, Chen et al. (2018) used a two branch network (one for each epoch of image data), where the feature maps of these streams were then merged by computing a Manhattan distance between them.

6.3 *Datasets and CNN input generation*

This section presents the image datasets used in the experiments and the process from the original aerial oblique images to the input given to the approaches indicated in section 4.

The datasets used in this chapter comprise two airborne oblique image captures of the city of L'Aquila and a smaller neighboring village, Tempera. These two locations were surveyed within an approximately 9-month interval, in August 2008 and in May 2009, the latter depicting the situation after the April 2009 earthquake that occurred in central Italy.

The image capture was performed using the Pictometry system that contains small format DSLR cameras, four obliques (one for each cardinal direction) and one nadir. The flying height was approximately 1000 m, which translated to an average sampling distance of 0.14 m on the oblique views. The flight was performed considering a forward overlap between 60 -70% and a side overlap between 35-45%.

Figure 44 depicts the process between the original images and the final input to the experiments. Two types of input were generated, façade image patches extracted from the original images (Figure 45, top), and these same image patches rectified using the corresponding façade 3D information (Figure 45, bottom). These were the two types of input that were used, and compared, in the experiments.

The first step was to generate the 3D point cloud, which was used to define the façade planes and subsequently extract the façades from the images. To this end the first step was to perform the image orientation of both pre- and post-event images. These shared the tie point computation with the objective of aligning the datasets. However, only the pre-event images were used for dense matching.

Figure 44 presents the overview of the main steps to extract the façade image patches from the oblique views using the 3D point cloud generated from the pre-event images. The first step was to differentiate between *on* and *off* ground points, using *lasground* from the package *lastools* (Axelsson, 2000). The point cloud, with the added attribute of the normalized height surface, was the input for a plane-based segmentation, which was followed by a connected component analysis, generating the final roof segments (Vosselman, 2012).

The roof segments were then projected into the *xy* plane (see Figure 44 – Façade definition). The approach then assumed that each building segment contains 4 façades and that they are mutually perpendicular. With this assumption the roof points were fitted with a minimum-area bounding rectangle (Freeman and Shapira, 1975) (red rectangle in Figure 44 - Façade definition) , defining the 4 main façade directions of a given building. The façade planes were then defined by the *xy* coordinates of the edges of the rectangle, where the *z* is obtained from the normalized height and from the difference between the mean *z* coordinate of the roof and the mean normalized height value. At this stage every façade was finally defined by 4 facade corners.

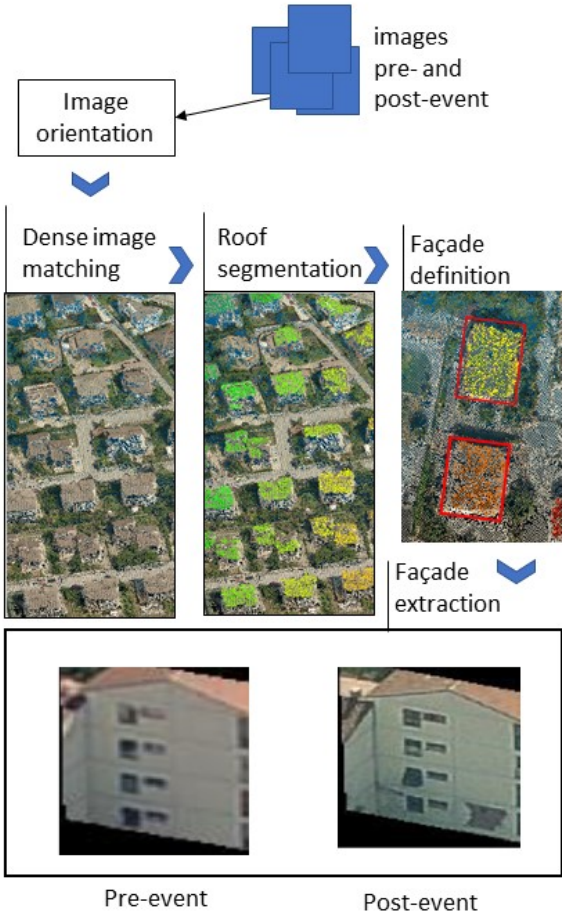

Figure 44 Overview of the main steps of the façade extraction from the aerial images. The segments in the Roof segmentation thumbnail are color coded. The red rectangle in the Facade definition thumbnail indicates the main 4 façades extracted from the roof points. Below, example of a façade, showing both pre- and post-event. These façade image patches (image pair) are one of the inputs to the experiments (see Figure 45).

The projection matrices, coming from the orientation step, were then used to project the facade pixels into these 3D planes. This process was repeated for all images containing a given façade, in both epochs. The same resolution was forced on the façades: gaps due to different viewpoints and scales of the oblique views were interpolated using a nearest neighbors' algorithm. This ensured the registration of the different views of the same façade. The visibility of the four façade corners in the images was used to detect occluded parts. The pre-event point cloud was used to perform the visibility analysis as it was assumed that all the buildings are still standing, and the number of occlusions is higher. A façade was considered occluded if at least two of the four corners were not visible in the image.

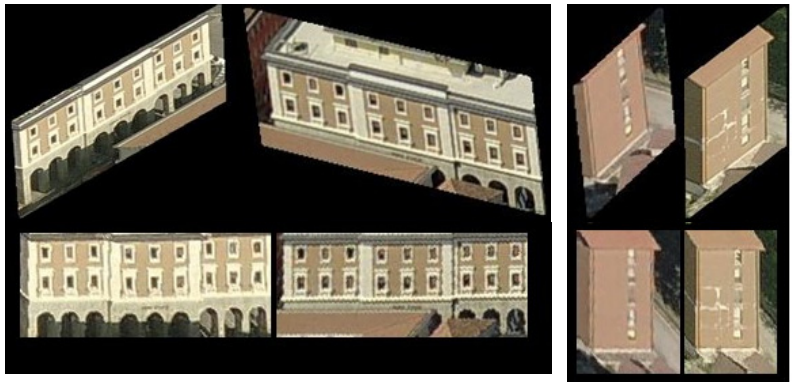

Figure 45 The two types of input used in the experiments, considering two views of two façades. Each of this pairs is an example of the input used in one set of experiments (see Figure 48). Top, original facade image patches. Bottom, corresponding rectified façade image patches.

Examples of the two types of extracted (original and rectified façade image patches) data can be seen in Figure 45. All the experiments, both mono- and multi-temporal, were tested considering separately the original and the rectified façade image patches. The aim was to test if the approaches could leverage the rectification and registration of the façade image patches to perform a better image classification of façade damages, while at the same time assuming the interpolated areas which might modify the already small damage evidences present in the façades.

To take advantage of having several façade image patches per façade per epoch, these pre- and post-event image data (original and rectified) were combined in two distinct ways:

1) Image pairs – these were created associating each pre-event façade image patch to all the post-event façade image patches of a given façade (Figure 2). This was performed for all façades. This input is related with the experiments MTa (see section 4.3). Performing this combination between different views of the same façade allowed to generate more input data, instead of considering an image patch per façade, which would make the training dataset very small (only 88 damaged and 90 undamaged façades were possible to extract from the images).
2) Image sextuples – these were created combining three pre-event image patches, with three post-event image patches of a given façade. In this case the maximum amount of combinations allowed per façade was 50, given an unbalanced number of views per facade. This input is related with the experiments MTb (see section 4.3). Considering several views per façade per epoch could enable the network to learn the similarities between different views of the same façade. This is due

to the fact that these networks compute features that are shared across all the different views, instead of focusing on single image pairs like in 1).

This allowed the extraction of 4,546 image pairs and 5,179 image sextuples from a total of 178 façades (see Table 12).

Table 12 Number of image pairs and image sextuples extracted considering the 178 façades.

	Image pairs	Image sextuples	Façades
Damaged	2,274	2,559	88
Not damaged	2,272	2,610	90
Total	4,546	5179	178

6.4 *Methodology*

Six multi-temporal approaches were designed, tested, and compared with three mono-temporal approaches. The multi-temporal approaches assumed as input pre- and post-event façade image patches captured from different oblique views (original and rectified), as described in the data section. The focus of the experiments was on the optimal merging of pre- and post-event image information within a supervised deep learning framework for the image classification of façade damages using CNN. Table 12 illustrates the small amount of data to perform this multi-temporal façade analysis using aerial manned imagery. This issue was central to the current work and is one of the main limitations of the experiments. Several measures were taken to attenuate the lack of data, and these are further detailed in this section and in the experiments.

This section starts by laying out the main characteristics of the used CNN, in the following paragraphs. The following sub-section formalizes the used network, while the final sub-sections explain the rationale behind each performed test.

6.4.1 Network definition

The basic network used in the experiments is presented in this sub-section, and it was a central component of the mono- and multi-temporal approaches (see *stream* in Figure 47,Figure 48 and Figure 49). This network was composed by consecutively stacking of 2 modules, dense blocks and transitional layers. This composition was proposed in (Huang et al., 2017), where the authors derived several networks from different combinations of these modules. In the current work, the used network was composed of 4 dense blocks, with transitional layers between these blocks. While maintaining the number of dense blocks presented in (Huang et al., 2017), a lower number of layers per dense block was considered in this work. Given the small amount of data,

decreasing the model complexity did not impact its representability and contributed to reduce overfitting.

Each dense block contained two sets of two convolutions, as indicated in Figure 46. In Figure 46, the conv field indicates the group: batch normalization, relu and convolution. A *dropout* layer (0.2) was also added after the first convolution, to further prevent overfitting (Clevert et al., n.d.). Each transitional layer contained a convolution and it was followed by average pooling with stride 2 in order decrease the feature map size from the initial 224x224px to the final 28x28px. The façade image patch given as input (both original and rectified) was zero padded to fit the 224x224px size. In rare cases where the façade image patch was larger than the 224x224px, it was resized to fit the input size while keeping the aspect ratio. This input size was mainly chosen to fit the fine-tuning experiment.

The number of filters per convolution was tied to the growth rate (Huang et al., 2017), which was defined as 6 (see Figure 46). This growth was set in order not to overfit the small set of data for the current study, while following the general assumption that more filters are needed to represent more complex features later in the network (Szegedy et al., 2014).

Given the small damage evidences often present in façades that did not collapse (see introduction figure) it was mandatory that a given network would be able to detect such small details. In this way it was important to retain contextual information, i.e. in the vicinity of the damaged area. Only then a network would be able to differentiate these small damage evidences from other areas with similar texture but in a different context. As can be seen in Figure 46, the dilation factor is also growing with the number of filters even if at a smaller rate. The last set of dilated convolutions had a receptive field of 19x19.

The classification part of the network was performed by coupling batch normalization, relu, global average pooling and a dense layer of size 1 at the end of all networks.

6.4.2 Mono-temporal approaches

Three mono-temporal approaches were tested. These served as baseline for the multi-temporal methods (see Figure 47).

The mono-temporal traditional (MN-trd) directly used a network trained with aerial image patches containing debris and rubble piles, as in (Duarte et al., 2018a). This network was trained on aerial (manned) image samples of 7 different geographical locations and using approximately 5,400 image samples in total. These locations include cities with similar urban design as to L'Aquila and Tempera (e.g., Amatrice, Italian city). For this approach each post-event façade image patch was divided into smaller 50px squared patches. The latter

were then classified for damage. In the case a façade image patch contained at least one of these squared patches classified as damage, the whole façade image patch was considered damaged. This was performed for every façade image patch of a given façade. This experiment aimed at understanding how a network trained solely on debris and rubble piles and mostly using nadir imagery could be used for the specific case of the detection of façade damages.

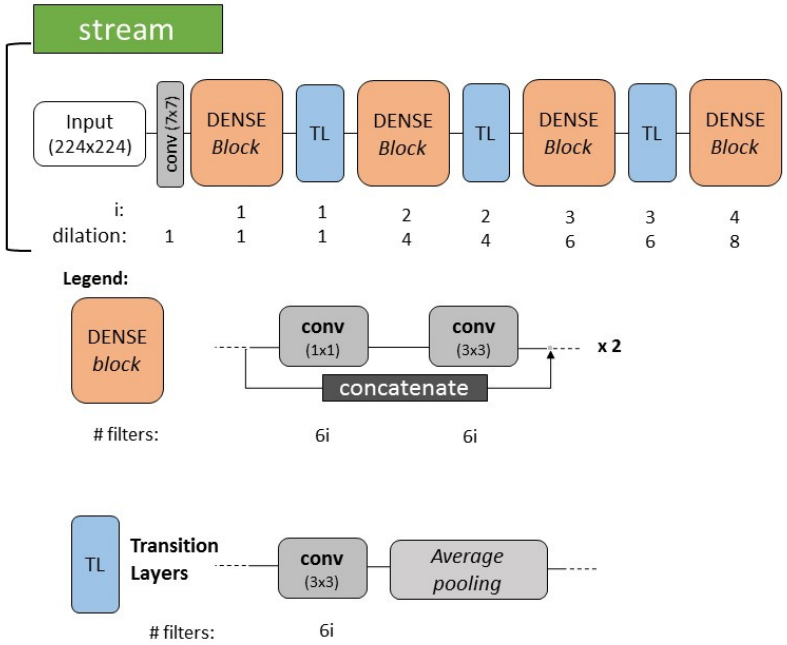

Figure 46 Network used in the experiments (stream), composed of dense blocks and transition layers. conv depicts the group batch normalization, relu and convolution. The number of filters and dilation value is affected by the number of dense block, transitional layer group, as indicated by i.

The other two mono-temporal approaches also only used post-event façade image patches. The mono-temporal, MN-scr, was trained from scratch, while the MN-ft was fine-tuned on *densenet* (DenseNet121 as in (Huang et al., 2017)), where only the last dense block layers were re-trained with the façade image patches coming from the different oblique views. This experiment was deemed necessary given the low amount of data and where the model could leverage the knowledge of the feature information learned on the ImageNET dataset.

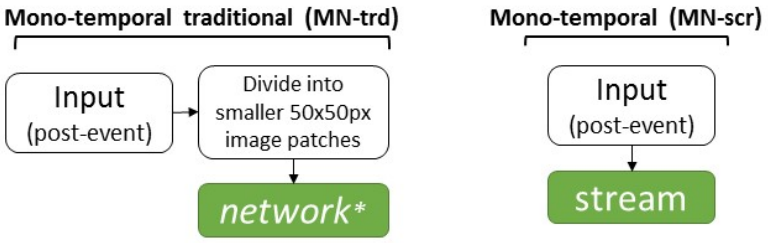

Figure 47 Mono-temporal approaches, MN-trd and MN-scr. * The network in italic refers to the aerial (manned) network presented in (Duarte et al., 2018a). The stream refers to the network presented in Figure 4. Input refers to façade image pairs.

6.4.3 Multi-temporal approaches

In this subsection, six multi-temporal approaches are presented. Overall, these experiments, aimed at: 1) better understanding how to merge the multi-temporal façade image patches within a CNN for the image classification of façade damages, and 2) embedding the façade image pairs and façade image sextuples defined in section 2.1 in the experiments. Figure 48 and Figure 49, show the six different approaches. Two different ways to integrate the data from multiple perspectives and epochs were adopted and tested in the approaches, respectively considering the image pairs (1) and image sextuples (2):

(1) The group MTa (see Figure 48) considered as input only image pairs, as described in section 3.1. In MTa three different strategies were adopted. The MTa-1str concatenated the images in the channels dimension and subsequently fed this to the network defined previously. On the other hand, MT-2str, assumed one convolutional block for each epoch which were later concatenated. The MT-2str-ws (or Siamese) was similar to MT-2str, but in this case the convolution weights were shared between the two streams.

(2) The group MTb (Figure 49) considered as input the image sextuples defined in section 3.1. The rationale of these approaches followed the same concept of the group MTa. The MTb-1str concatenated the six images (three per epoch), where this 18-channel image was then fed to the network, while MTb-2str considered a convolutional set per epoch. In this case a concatenation of the three images per epoch was performed, where this 9-channel image was fed to an independent convolutional block. MT-2str-sw (or Siamese) was only different from MTb-2str, given that the convolutions were shared across streams. In this case features were computed not only across epochs but also across different views, given that for each epoch several image façade patches per façade were simultaneously considered.

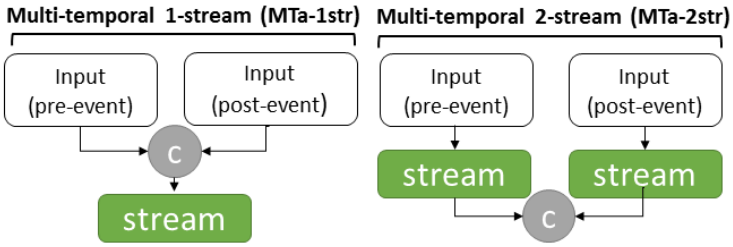

Figure 48 MTa group of experiments. Façade image pairs are fed to the experiments present in this figure.

The 1str set of experiments, both in MTa and MTb, forced the input image patches (both pre- and post-event) to go through a single convolutional set. On the other hand, 2str experiments had a set of epoch specific convolutions, where this information was later merged through concatenation. The 2str-sw (or Siamese) had the convolution weights shared across the epochs. These 3 different ways of considering the input data aimed at understanding which set of features were relevant for the image classification of façade damages. While the 1str approaches made use of inter-epoch features given that the inputs were concatenated at an early stage of the network, the 2str approaches gave more relevance to intra-epoch features which were merged at a later stage of the network. The 2str and 2str-sw only differed in the fact that the convolutions were shared across the epochs: in spite of having a set of convolutions for each epoch, these had the filters shared among them. Moreover, given the sharing of filters between epochs, this drastically decreased the number of parameters when compared with the 2str which did not share the convolutions.

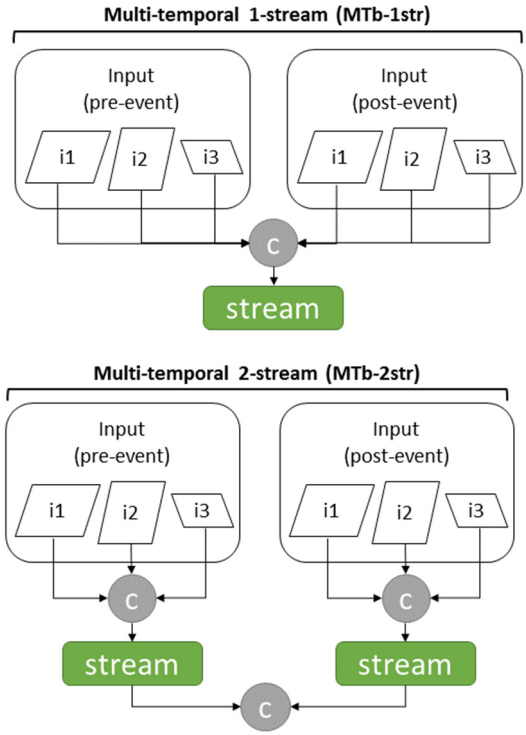

Figure 49 MTb group of experiments. Façade image sextuples are considered as input and indicated by i1-3 for each epoch.

While concatenation was used to merge the feature maps, other approaches were tested (e.g., element-wise multiplication or addition/subtraction of the feature maps). However, these did not perform as well as the simple concatenation.

All these approaches were tested using both the original and rectified façade images patches as described in 2.1.

6.5 *Experiments and Results*

All the networks were trained with learning rate of 0.1 and weight decay of 10^{-4} (except for the fine tune experiment, where the learning rate was of 0.01) using stochastic gradient descent as optimizer (He et al., 2016; Huang et al., 2017). For each experiment one loss function, binary cross entropy, was used, given the binary classification problem being considered. The experiments were performed with early stopping, i.e. when the validation data loss stopped improving. This was performed to avoid overfitting given the small data set.

Data augmentation was performed also in a bid to decrease overfitting and to give more generalization capabilities to the network (Krizhevsky et al., 2012). However, the data augmentation consisted only of horizontal shifts and image

normalization. This was due to two reasons: 1) shifts in the images could mask out the damaged area when it is close to the edge of the image; 2) rotations on the image patches could be cue for damage (e.g., slanted buildings which did not collapse). Given the small amount of data, this solution attenuated overfitting and helped generalization (Krizhevsky et al., 2012).

The results were evaluated in terms of accuracy, recall, precision and f1 score (as indicated in the equations 1, 2, 3 and 4). These were computed three times for each experiment. In each run of the experiment the data were randomly divided in sets of training and validation (70% and 30%, respectively). This division was performed at a façade level, where both the image pairs and image sextuples datasets are relative to the same façades and hence comparable. In this way every façade was present in both training and validation when considering the three different splits. The mean and the range (min. and max.) of the different runs per experiment are shown in the results, too.

$$accuracy = \frac{TP+FN}{\#\ validation\ samples} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$f1 = 2\frac{recall*accuracy}{recall+accuracy} \quad (4)$$

Overall, the multi-temporal approaches clearly outperformed the mono-temporal ones. This is seen in both in the MTa and MTb experiments, and also when using either the original façade image patches or the rectified ones.

In general, using an epoch-specific set of convolutions per epoch was preferable in all the multi-temporal experiments. However, the results differ when considering different inputs, original or rectified façade image patches. While having similar results, the MTb-2str-sw-r was the best performing approach when compared with MTa-2str. Hence, the use of shared convolutions (in a Siamese setting) is most valuable when considering the image sextuples using as input the rectified façade image patches. On the other hand, when using the original façade image patches, the network cannot take advantage of the simultaneous consideration of several views per façade.

The results of MTa-2str and MTb-2str-sw-r were considerably different when compared at an image pair/sextuple level, where the difference was bridged when evaluated at a façade level. While having less correctly predicted image pairs/sextuples, MTb-2str-sw-r (82% f1-score) outperformed MTa-2str (80% f1-score) at a façade level. Hence, the better results at an image pair/sextuple by MTa-2str were more distributed among the façades, not being enough to change the prediction at a façade level. On the other hand, MTb-2str-sw-r,

improved the results at a façade level while having less correctly predicted image pairs/sextuples. Hence, in this case the correct predictions were more distributed among the façades, which in turn, through the majority vote, improved the results at this level.

The overall statistical measures range between the three different data splits was also smaller when using the rectified image patches. In some of the experiments (e.g., MTa-2str-r and MTa-2srt-sw-r) recall and precision achieve 1.0 at least in one of the data splits, where the approach struggled to differentiate between the two classes. However, their non-rectified counterpart did not present this behavior, indicating that the combined use of rectified façade image patches and image pairs may not be optimal.

The mono-temporal approaches presented the worst results. The traditional approach trained on rubble piles and debris was the worst performing approach, especially using rectified façade image patches.

Figure 50 presents activations (right) considering a given façade (pre- and post-rectified façade image patches) (left). These activations were extracted from the last set of activations of each of the experiments predicting on an image sample that was present in training. This aimed at understanding where the approaches were focusing their attention on a given façade image patch, to derive a given class prediction. Figure 50 B, D and E were predicted as damaged. The only clear activation focusing on the damaged area is present in E. In the B and D cases, in spite of also considering the correct damaged area, these are not so clear and often consider other areas of the image. For example in D, post-event, the balconies area was relevant for the approach to derive the damaged class (also close to the damaged portion of the façade, see red indication in Figure 50) . A and C present damaged areas which were predicted as not damaged. In both cases the attention of the network was on almost the whole extent of the façades, detecting neither the small cracks in A, nor the small collapsed segment in C.

Table 13 Precision, recall, accuracy and f1 score (mean) for the mono- and multi-temporal approaches using the original façade image patches (range between brackets). These are presented at both an image pair/sextuple and at a façade level

	Image/image-pair/image-sextuple level			
	Prec.	Rec.	Acc.	F1
MN-trd	0.49 (0.48-0.58)	0.66 (0.55-0.73)	0.50 (0.47-0.55)	0.55 (0.52-0.64)
MN-scr.	0.58 (0.45-0.63)	0.85 (0.65-1.00)	0.64 (0.61-0.68)	0.72 (0.53-0.73)
MN-ft	0.64 (0.57-0.64)	0.84 (0.47-0.96)	0.63 (0.57-0.64)	0.67 (0.54-0.76)
MTa-1str	0.73 (0.65-0.77)	0.60 (0.60-0.69)	0.72 (0.64-0.74)	0.67 (0.62-0.71)
MTa-2str	0.83 (0.79-0.85)	0.76 (0.57-0.80)	0.81 (0.72-0.83)	0.80 (0.66-0.82)
MTa-2str-ws	0.76 (0.66-0.81)	0.60 (0.56-0.89)	0.73 (0.65-0.83)	0.69 (0.61-0.82)
MTb-1str	0.76 (0.66-0.78)	0.64 (0.52-0.64)	0.73 (0.64-0.80)	0.70 (0.58-0.81)
MTb-2str	0.71 (0.66-0.77)	0.64 (0.52-0.87)	0.76 (0.64-0.71)	0.70 (0.58-0.70)
MTb-2str-sw	0.77 (0.64-0.77)	0.75 (0.55-0.78)	0.76 (0.63-0.78)	0.75 (0.59-0.78)
	Façade level			
MN-trd	0.52 (0.43-0.63)	0.67 (0.38-0.70)	0.55 (0.52-0.60)	0.60 (0.40-0.65)
MN-scr.	0.55 (0.50-0.60)	0.92 (0.67-1.00)	0.67 (0.52-0.74)	0.70 (0.57-0.72)
MN-ft	0.65 (0.54-0.70)	0.61 (0.38-0.67)	0.63 (0.59-0.70)	0.60 (0.49-0.63)
MTa-1str	0.69 (0.67-0.9)	0.82 (0.56-0.83)	0.72 (0.62-0.87)	0.74 (0.62-0.86)
MTa-2str	0.88 (0.88-0.92)	0.73 (0.64-0.79)	0.82 (0.80-0.86)	0.80 (0.74-0.85)
MTa-2str-ws	0.75 (0.70-0.81)	0.58 (0.38-0.81)	0.68 (0.59-0.83)	0.64 (0.51-0.81)
MTb-1str	0.78 (0.67-0.81)	0.58 (0.46-0.93)	0.70 (0.58-0.86)	0.67 (0.55-0.87)
MTb-2str	0.72 (0.67-0.80)	0.73 (0.56-0.73)	0.78 (0.59-0.78)	0.76 (0.56-0.76)
MTb-2str-sw	0.82 (0.75-0.85)	0.73 (0.50-0.82)	0.79 (0.60-0.82)	0.79 (0.67-0.83)

In Figure 51 several correct (on the left) and wrong (on the right) predictions are shown considering the best performing approach. This approach correctly identifies several degrees of spalling, building segment collapses and larger cracks. However, it is not able to detect areas with small spalling when these are too small when compared with the size of the façade. Façades which only presented cracks were often missed by the approach, probably because of the limited resolution of the images.

Table 14 Precision, recall, accuracy and f1 score (mean) for the mono- and multi-temporal approaches using the rectified (-r) façade image patches (range between brackets). These are presented at both an image pair/sextuple and at a façade level

	Image/image-pair/image-sextuple level			
	Prec.	Rec.	Acc.	F1
MN-trd-r	0.39 (0.34-0.80)	0.37 (0.26-0.38)	0.50 (0.47-0.64)	0.38 (0.30-0.52)
MN-scr.-r	0.65 (0.48-0.72)	0.83 (0.71-0.88)	0.70 (0.65-0.72)	0.71 (0.64-0.77)
MN-ft-r	0.69 (0.68-0.94)	0.68 (0.22-0.68)	0.69 (0.59-0.70)	0.68 (0.36-0.69)
MTa-1str-r	0.76 (0.70-0.80)	0.67 (0.52-0.77)	0.73 (0.67-0.76)	0.73 (0.62-0.73)
MTa-2str-r	0.70 (0.66-0.88)	0.72 (0.65-0.94)	0.73 (0.70-0.79)	0.78 (0.68-0.79)
MTa-2str-ws-r	0.86 (0.63-0.87)	0.64 (0.53-0.81)	0.73 (0.68-0.77)	0.71 (0.66-0.73)
MTb-1str-r	0.69 (0.67-0.7)	0.86 (0.72-0.96)	0.74 (0.71-0.76)	0.77 (0.71-0.79)
MTb-2str-r	0.71 (0.64-0.71)	0.64 (0.61-0.64)	0.69 (0.65-0.69)	0.66 (0.64-0.66)
MTb-2str-sw-r	0.76 (0.72-0.89)	0.75 (0.67-0.83)	0.76 (0.71-0.79)	0.75 (0.62-0.81)
	Façade level			
MN-trd-r	0.47 (0.46-0.56)	0.47 (0.32-0.58)	0.58 (0.52-0.66)	0.51 (0.38-0.52)
MN-scr.-r	0.65 (0.53-0.66)	0.92 (0.62-1.00)	0.69 (0.69-0.70)	0.69 (0.64-0.76)
MN-ft-r	0.66 (0.66-1.00)	0.62 (0.22-0.62)	0.69 (0.56-0.70)	0.64 (0.30-0.64)
MTa-1str-r	0.80 (0.80-0.89)	0.66 (0.57-0.73)	0.74 (0.72-0.84)	0.72 (0.67-0.80)
MTa-2str-r	0.67 (0.67-1.00)	0.73 (0.67-1.00)	0.76 (0.72-0.83)	0.80 (0.70-0.80)
MTa-2str-ws-r	0.78 (0.60-1.00)	0.58 (0.67-1.00)	0.68 (0.66-0.70)	0.67 (0.55-0.70)
MTb-1str-r	0.69 (0.68-0.71)	0.84 (0.83-0.93)	0.74 (0.71-0.76)	0.77 (0.76-0.79)
MTb-2str-r	0.71 (0.67-0.83)	0.71 (0.55-0.77)	0.71 (0.65-0.79)	0.74 (0.60-0.77)
MTb-2str-sw-r	0.87 (0.76-0.94)	0.80 (0.45-0.95)	0.84 (0.76-0.85)	0.82 (0.62-0.87)

6.6 *Discussion*

All the multi-temporal approaches outperformed the mono-temporal ones. This confirms the commonly reported results on multi-temporal approaches in remote sensing, where the use of multi-temporal data is often translated into an improvement in the quality of a given task (Hussain et al., 2013; Lu et al., 2004; Singh, 1989; Tewkesbury et al., 2015). The best performing approach can identify partially and totally collapsed buildings, and façades with large areas of spalling and cracks. However, the overall results (best performing network with 82% f1-score) reflect some difficulties in the detection of damage to the facades, from manned aerial oblique imagery, even when also using pre-event images. This can be mainly explained by the low resolution of the used data (GSD ~0.14 m) that hinders the reliable detection of smaller signs of damage.

From all the mono-temporal approaches, fine-tuning or training from scratch did not lead to considerable differences as reported in other works focused on building damage detection (Duarte et al., 2018a; Vetrivel et al., 2017). The

mono-temporal traditional approach using a model trained with image samples depicting rubble piles and debris was the worst approach: as expected the model was not able to identify lower levels of damage present in the façades. This resulted in a high rate of both false negatives and false positives as in Duarte et al. (2017). This problem was more accentuated when using the rectified façade image patches, as the traditional approach was trained on non-rectified image patches.

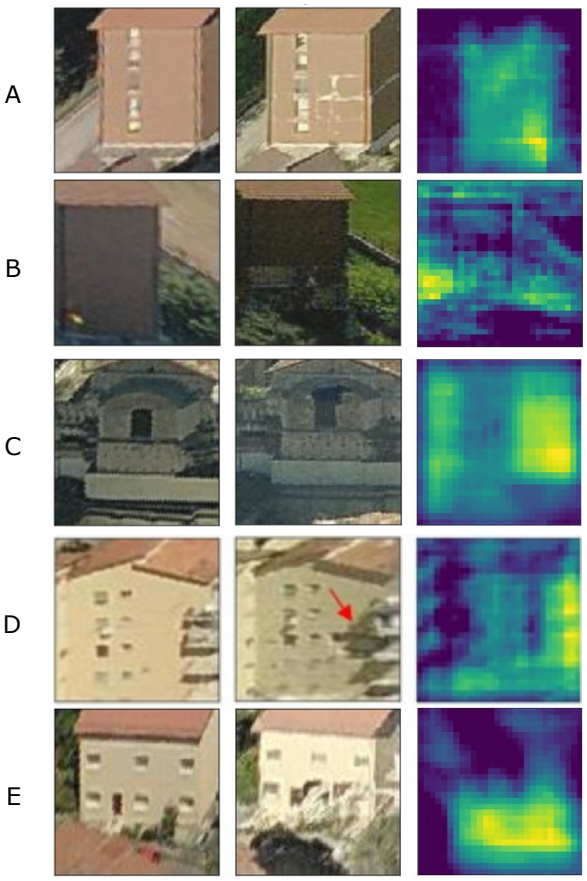

Figure 50 Activations extracted from the last activation layer of the network (training) MTb-2str-sw-r (right). Left(pre-event) and middle (post-event) facade image patches. A, C predicted as not damaged, while B, D and E were predicted as damaged.

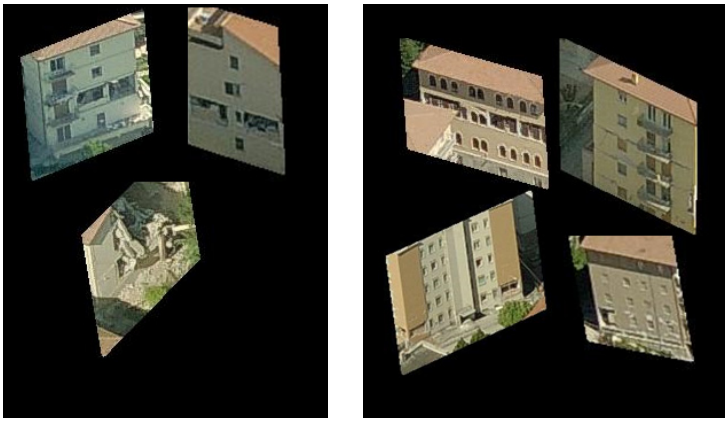

Figure 51 Left, correctly classified as damaged. Right, incorrectly classified as not-damage. Both using the best performing approach MTb-2str-sw-r, when these façades were not present in training.

Regarding the multi-temporal approaches the relevance of intra-epoch features must be noticed, which are merged later in the network. This can be observed in the results, where the single stream approaches were always outperformed by the 2str approaches, independently of the use of original or rectified image patches or the input data (i.e. image pairs or sextuples). Recent literature in remote sensing that made use of multi-branch networks reported different results in this regard. For example, Daudt et al. (2018) reported that for the specific case of satellite imagery change detection the concatenation of the images before being fed to the network would be preferable, as images share the features within a single convolutional set. This was also the case when localizing street view images using overhead images (Vo and Hays, 2016). However, there are also studies in which the merging of the feature information later in the network, instead of considering as input a merged layer of both epochs, is preferable (Chen et al., 2018). The merging of the pre- and post-event information seems to be application dependent, where for the case of the image classification of façade damages it is preferred to merge the feature information at a later stage in the network. Also, the differences between the results at an image pair/sextuple and at a façade level are noteworthy. Since in most of the approaches the façade level results were worse than its image pair/sextuple counterpart, it seems that in such situations there was no considerable variation of the predictions within the same façade.

In the case where the façade image patches are rectified, the approach using the sextuples as input outperforms all the other approaches (MTb-2str-sw-r) at a façade level. Besides using image sextuples and rectified façade image patches the approach also shared the convolutions between the two streams. Given the rectification and registration of the façades, this approach aimed at taking advantage of considering different viewpoints of the same façade simultaneously. In this way it was expected that the networks would leverage

the multi-view information, extracting features across both the different epochs and the different views. However, at a façade level the approach considering only image pairs (MTa-2str) performed comparatively well (difference of 2% f1-score) while using the original image patches. In this regard, although the rectification/registration procedure is preferable, it may at the same time smooth out the often small damage evidences present in the façades.

In this study only 88 damaged façades were extracted, where the high overlap of manned aerial systems allowed to derive several image pairs and image sextuples per façade, per epoch. In this way it was possible to perform the experiments reported in this chapter. This is an understudied subject, where usually the redundancy of aerial surveys is not fully used.

Although the image coverage of the area is relatively high, another limitation was given by the occlusions in dense urban areas: several buildings or part of them were almost invisible in the images. This is an intrinsic limitation of aerial-manned platforms with pre-defined flight patterns not tailored to decrease such occlusions. In this sense more careful flight plans and using UAV could attenuate this problem.

6.7 *Conclusions*

This chapter assessed the image classification of façade damages using multi-temporal aerial oblique imagery. Six multi-temporal and three mono-temporal approaches were tested, following a binary classification approach using CNN. For this purpose, the only dataset (to the best of the authors' knowledge) available with pre- and post-event data was used for this analysis. Although the dataset is not optimal in terms of number of images and resolution, it has shown very encouraging results and good indications for the wide adoption of multi-temporal data in the assessment of catastrophic event damages.

The objective of this study was twofold: 1) determine the optimal framework to combine the multi-temporal image data within a CNN approach, and 2) investigate the improvement introduced by the use of the multi-view characteristics of aerial (manned) systems (extracting several image patches per façade and per epoch) in the image classification of façade damages. In this regard two main approaches were tested: 1) using image pairs by pairing every pre-event façade image patch to the corresponding post-event façade image patches, and 2) using image sextuples where three views per façade per epoch were considered.

An important element tested in this chapter was the use of rectified façades instead of the original façade image patches. Regarding the original façade image patches, the best approach was to use image pairs and a 2-stream network (no shared convolutions) while using rectified façade image patches,

the use of the image sextuples and shared convolutions was more advantageous. Given the rectification and registration of the façade image patches, considering three views per epoch only slightly improved over the approach considering image pairs. A study considering more data would need to be performed to assess if the network can learn not only inter epoch dependencies, but also to cope with different views of a given façade.

The multi-temporal approaches generally outperformed the mono-temporal ones. Large differences in the multi-temporal results were, however, visible according to the used network. The use of epoch-specific convolutions was preferable to single stream architectures, where both epochs inputs are concatenated together before being fed to the network. Epoch-specific feature information is in this way valuable for the image classification of façade damages. This was the case regardless of the use of original or rectified image patches as input, and regardless of the use of image pairs or image sextuples. However, while the best performing network using the original image pairs considered a 2-stream network without shared convolutions, this was not the case when using the rectified façade image patches where the 2-stream network sharing the convolutions (i.e. Siamese) was preferable.

Regarding the mono-temporal approaches, the network trained on image samples depicting debris and rubble piles was often not able to have a better score than random guess (i.e. 50% accuracy). Hence, such networks trained with only rubble piles and debris from mostly nadir viewing imagery, are not transferable for façade cases where damage evidences are often different in image content but also in extent (e.g., small signs of spalling or cracks). The mono-temporal approach using damaged and non-damaged façade image patches performed better when trained from scratch; however, overall it behaved poorly.

A notable limitation of the approach presented here is its binary nature that precludes more nuanced damage assessment. In the disaster response phase, the location of partially and totally collapsed buildings is a priority. Hence, in such case the binary nature of the approach is not sufficient, since it considers several typologies of damage (from spalling to completely destroyed façades). More work is needed, based on more oblique multi-temporal image datasets, to move towards the classification of the different types of façade damages and their localization within the façade. Nonetheless, given the focus of this work on the specific façade damage detection, this approach could be performed in parallel with the already extensively reported methods in the literature to detect rubble piles and debris.

The used dataset was extremely challenging not only for the limited number of images (and facades) and the low resolution, but for the urban typology (historical city center) that introduced additional challenges. Several façades in the test area were impossible to extract given the often narrow streets. This

was further exacerbated by the use of an aerial (manned) system and its pre-defined flight pattern, which limited the data completeness in some narrow streets.

6.8 *References of Chapter 6*

Armesto-González, J., Riveiro-Rodríguez, B., González-Aguilera, D., Rivas-Brea, M.T., 2010. Terrestrial laser scanning intensity data applied to damage detection for historical buildings. J. Archaeol. Sci. 37, 3037–3047. https://doi.org/10.1016/j.jas.2010.06.031

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011

Audebert, N., Le Saux, B., Lefèvre, S., 2017. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks, in: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (Eds.), Computer Vision – ACCV 2016. Springer International Publishing, Cham, pp. 180–196. https://doi.org/10.1007/978-3-319-54181-5_12

Axelsson, P., 2000. DEM generation from laser scanning data using adaptive TIN models, in: International Archives of Photogrammetry and Remote Sensing. Presented at the XIX ISPRS Congress, Amsterdam, pp. 111–118.

Brunner, D., Schulz, K., Brehm, T., 2011. Building damage assessment in decimeter resolution SAR imagery: A future perspective, in: Joint Urban Remote Sensing Event. IEEE, pp. 217–220. https://doi.org/10.1109/JURSE.2011.5764759

Chen, Z., Zhang, Y., Ouyang, C., Zhang, F., Ma, J., 2018. Automated landslides detection for mountain cities using multi-temporal remote sensing imagery. Sensors 18, 821. https://doi.org/10.3390/s18030821

Clevert, D.-A., Unterthiner, T., Hochreiter, S., n.d. Fast and accurate deep network learning by exponential linear units (ELUs), in: ICLR 2016. Presented at the ICLR 2016.

Curtis, A., Fagan, W.F., 2013. Capturing damage assessment with a spatial video: an example of a building and street-scale analysis of tornado-related mortality in Joplin, Missouri, 2011. Ann. Assoc. Am. Geogr. 103, 1522–1538. https://doi.org/10.1080/00045608.2013.784098

Cusicanqui, J., Kerle, N., Nex, F., 2018. Usability of aerial video footage for 3D-scene reconstruction and structural damage assessment. Nat. Hazards Earth Syst. Sci. Discuss. 1–23. https://doi.org/10.5194/nhess-2017-409

Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y., 2018. Urban change detection for multispectral earth observation using convolutional neural networks. Presented at the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Valencia, pp. 2115–2118. https://doi.org/10.1109/IGARSS.2018.8518015

Dell'Acqua, F., Gamba, P., 2012. Remote sensing and earthquake damage assessment: experiences, limits, and perspectives. Proc. IEEE 100, 2876–2890. https://doi.org/10.1109/JPROC.2012.2196404

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS J. Photogramm. Remote Sens. 84, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2019. Damage detection on building façades using multi-temporal aerial oblique imagery. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-2/W5, 29–36. https://doi.org/10.5194/isprs-annals-IV-2-W5-29-2019

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2018a. Multi-resolution feature fusion for image classification of building damages with convolutional neural networks. Remote Sens. 10, 1636. https://doi.org/10.3390/rs10101636

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2018b. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach, in: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 89–96. https://doi.org/10.5194/isprs-annals-IV-2-89-2018

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2017. Towards a more efficient detection of earthquake induced facade damages using oblique UAV imagery, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 93–100. https://doi.org/10.5194/isprs-archives-XLII-2-W6-93-2017

Dubois, D., Lepage, R., 2014. Fast and efficient evaluation of building damage from very high resolution optical satellite images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 4167–4176. https://doi.org/10.1109/JSTARS.2014.2336236

Fernandez Galarreta, J., Kerle, N., Gerke, M., 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Nat. Hazards Earth Syst. Sci. 15, 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

Freeman, H., Shapira, R., 1975. Determining the minimum-area encasing rectangle for an arbitrary closed curve. Commun. ACM 18, 409–413. https://doi.org/10.1145/360881.360919

Gerke, M., Kerle, N., 2011. Automatic structural seismic damage assessment with airborne oblique Pictometry© imagery. Photogramm. Eng. Remote Sens. 77, 885–898. https://doi.org/10.14358/PERS.77.9.885

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., Hikosaka, S., 2017. Effective use of dilated convolutions for segmenting small object instances in remote sensing images arXiv:1709.00179.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR. https://doi.org/10.1109/CVPR.2016.90

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7, 14680–14707. https://doi.org/10.3390/rs71114680

Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H., 2015. Deep Convolutional Neural Networks for Hyperspectral Image Classification. J. Sens. 2015, 1–12. https://doi.org/10.1155/2015/258619

Huang, G., Liu, Z., Maaten, L. van der, Weinberger, K.Q., 2017. Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS J. Photogramm. Remote Sens. 80, 91–106. https://doi.org/10.1016/j.isprsjprs.2013.03.006

Jiang, H., Lu, N., 2018. Multi-scale residual convolutional neural network for haze removal of remote sensing images. Remote Sens. 10, 945. https://doi.org/10.3390/rs10060945

Jung, J., Yun, S.-H., Kim, D., Lavalle, M., 2018. Damage-mapping algorithm based on Coherence model using multitemporal polarimetric–interferometric SAR data. IEEE Trans. Geosci. Remote Sens. 56, 1520–1532. https://doi.org/10.1109/TGRS.2017.2764748

Kampffmeyer, M., Salberg, A., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban Remote Sensing images using deep convolutional neural networks. Presented at the CVPR 2016.

Kerle, N., Hoffman, R.R., 2013. Collaborative damage mapping for emergency response: the role of Cognitive Systems Engineering. Nat. Hazards Earth Syst. Sci. 13, 97–113. https://doi.org/10.5194/nhess-13-97-2013

Khoshelham, K., Oude Elberink, S., Sudan Xu, 2013. Segment-based classification of damaged building roofs in aerial laser scanning data. IEEE Geosci. Remote Sens. Lett. 10, 1258–1262. https://doi.org/10.1109/LGRS.2013.2257676

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1907–1105.

Längkvist, M., Kiselev, A., Alirezaie, M., Loutfi, A., 2016. Classification and segmentation of satellite orthoimagery using convolutional neural networks. Remote Sens. 8, 329. https://doi.org/10.3390/rs8040329

Li, X., Chen, X., Liang, L., Xiao Chen, Lei Liang, 2012. A new approach to collapsed Building extraction using RADARSAT-2 polarimetric SAR imagery. IEEE Geosci. Remote Sens. Lett. 9, 677–681. https://doi.org/10.1109/LGRS.2011.2178392

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: CVPR. IEEE, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Lu, D., Mausel, P., Brondízio, E., Moran, E., 2004. Change detection techniques. Int. J. Remote Sens. 25, 2365–2401. https://doi.org/10.1080/0143116031000139863

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657. https://doi.org/10.1109/TGRS.2016.2612821

Nex, F., Duarte, D., Steenbeek, A., Kerle, N., 2019. Towards Real-Time Building Damage Mapping with Low-Cost UAV Solutions. Remote Sens. 11, 287. https://doi.org/10.3390/rs11030287

Nogueira, K., Penatti, O.A.B., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognit. 61, 539–556. https://doi.org/10.1016/j.patcog.2016.07.001

Persello, C., Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in VHR images. IEEE Geosci. Remote Sens. Lett. 14, 2325–2329. https://doi.org/10.1109/LGRS.2017.2763738

Singh, A., 1989. Review Article Digital change detection techniques using remotely-sensed data. Int. J. Remote Sens. 10, 989–1003. https://doi.org/10.1080/01431168908903939

Sui, H., Tu, J., Song, Z., Chen, G., Li, Q., 2014. A novel 3D building damage detection method using multiple overlapping UAV images. ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XL–7, 173–179. https://doi.org/10.5194/isprsarchives-XL-7-173-2014

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions arXiv:1409.4842.

Tewkesbury, A.P., Comber, A.J., Tate, N.J., Lamb, A., Fisher, P.F., 2015. A critical synthesis of remotely sensed optical image change detection techniques. Remote Sens. Environ. 160, 1–14. https://doi.org/10.1016/j.rse.2015.01.006

Tu, J., Sui, H., Feng, W., Sun, K., Xu, C., Han, Q., 2017. Detecting building façade damage from oblique aerial images using local symmetry feature and the Gini Index. Remote Sens. Lett. 8, 676–685. https://doi.org/10.1080/2150704X.2017.1312027

United Nations, 2015. INSARAG guidelines, volume II: preparedness and response, manual B: operations.

United Nations, 2009. 2009 UNISDR Terminology on disaster risk reduction. United Nations International Strategy for Disaster Reduction, Geneva, Switzerland.

Vetrivel, A., Duarte, D., Nex, F., Gerke, M., Kerle, N., Vosselman, G., 2016. Potential of multi-temporal oblique airborne imagery for structural damage assessment. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. III–3, 355–362. https://doi.org/10.5194/isprsannals-III-3-355-2016

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2017. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Vo, N.N., Hays, J., 2016. Localizing and orienting street views using overhead imagery, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. Springer International Publishing, Amsterdam, pp. 494–509. https://doi.org/10.1007/978-3-319-46448-0_30

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55, 881–893. https://doi.org/10.1109/TGRS.2016.2616585

Vosselman, G., 2012. Automated planimetric quality control in high accuracy airborne laser scanning surveys. ISPRS J. Photogramm. Remote Sens. 74, 90–100. https://doi.org/10.1016/j.isprsjprs.2012.09.002

Wallemacq, P., House, R., 2018. Economic losses, poverty & disasters: 1998-2017.

Wang, Q., Zhang, X., Chen, G., Dai, F., Gong, Y., Zhu, K., 2018. Change detection based on Faster R-CNN for high-resolution remote sensing images. Remote Sens. Lett. 9, 923–932. https://doi.org/10.1080/2150704X.2018.1492172

Xia, G.-S., Wang, Z., Xiong, C., Zhang, L., 2015. Accurate annotation of remote sensing images via active spectral clustering with little expert knowledge. Remote Sens. 7, 15014–15045. https://doi.org/10.3390/rs71115014

Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks, in: CVPR.

Zhang, C., Wei, S., Ji, S., Lu, M., 2019. Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-nased classification. ISPRS Int. J. Geo-Inf. 8, 189. https://doi.org/10.3390/ijgi8040189

Zhuo, X., Fraundorfer, F., Kurz, F., Reinartz, P., 2019. Automatic annotation of airborne images by label propagation based on a bayesian-CRF model. Remote Sens. 11, 145. https://doi.org/10.3390/rs11020145

7 Synthesis

The research presented in this dissertation focused on the development of methods for the identification of rubble piles, debris, and façade damages from remote sensing images. The identification of such damaged areas is of utmost importance for several stages of the disaster management cycle. The identification of rubble piles and debris from remote sensing images may serve for the identification of partially and totally collapsed buildings over a region, city or even a building block. Such information is critical in the response phase, so first responders (FR) can plan their rescue efforts, where the generation of damaged maps needs to be both fast and accurate to be of use. Spalling, cracks and other signs of façade damage as well as a per building segment damage assessment is central for the recovery and rehabilitation phase, given its contribution to the broader task of a per building damage assessment. The approaches presented in this dissertation used image data that were captured from three platforms: satellite and aerial (manned and unmanned), for the mapping of building damages. In the specific case of façade damages only aerial oblique views were used, since these directly survey the façades. The findings and results of the approaches are presented and summarized below. The research is also contextualized within developments which occurred during its execution and also within the INACHUS project.

The mapping of partially and totally collapsed buildings relies on the identification of debris and rubble piles from remote sensing imagery. Mapping such damage evidences is often performed specifically considering the used system (platform and sensor) (Dubois and Lepage, 2014; Fernandez Galarreta et al., 2015; Vetrivel et al., 2017). In the case the objective is to identify rubble piles and debris from satellite images, features are extracted from these images and used to identify damages on new image patches (Dubois and Lepage, 2014). However, such task is constrained by the amount of available imagery (captured from a given platform) to extract these features. This is more critical considering the state-of-the-art in image classification approaches such as convolutional neural networks (CNN) which need large amounts of image data for the classifiers to have image recognition capabilities. The experiments presented in sections 2 and 3 address this issue by proposing a unified damage detection procedure which makes use of all the imagery containing rubble piles and debris, regardless of the platform which was used for its capture. The first set of experiments focused on the use of aerial (manned and UAV) and satellite image samples for the image classification of debris and rubble piles in satellite images following a binary, patch-based CNN. Two main approaches were considered: 1) focusing on the sharing of features between the different resolutions within a single set of convolutions, 2) having a resolution specific set of convolutions for each of the resolution levels. The latter was preferred, where the satellite image classification of building damages was improved almost 4% when comparing with a classical approach (mono-resolution trained only with satellite image patches). The results indicate that stronger image classification algorithms for the mapping of debris

and rubble piles can be generated when extracting features also from aerial images. Up to now this was not the case, where such images would not be considered at all for the specific case of image classification of building damages from satellite images. Moreover, it was observed that the activation maps from the last set of convolutions had coarse localization capabilities even if a patch-based method was used. Such information can be used to generate heat maps of building damages within the image patches being classified

Given the results obtained before when using satellite images, the multi-resolution approach was then extended to the other resolutions, aerial manned and UAV. Chapter 3 presents the experiments for each of the resolution levels, satellite, aerial manned, and UAV, when using a multi-resolution approach. As happened before when only focusing on the satellite case, in this case to merge the epoch specific convolutional sets was also preferable (~3% difference in accuracy). However, results varied between resolutions. While in the case of satellite and UAV, the use of a multi-resolution approach improved their image classification accuracy; this was not the case for the aerial case in which there was no improvement. Aerial manned datasets are usually captured with high-end calibrated cameras and with a more homogenous data capture when compared with satellite and the UAV datasets. Hence, using image data from other resolutions only matched the image classification accuracy of mono-resolution approaches in the aerial manned case. The impact of a multi-resolution approach was then tested for geographical transferability in order to assess its impact when used on unseen locations with different urban design and image capture characteristics, namely in the UAV and satellite case. Overall the relative differences between the baseline and the multi-resolution experiments were maintained, especially for the satellite and UAV case. Nonetheless, all the geographical transferability experiments suffered from a decrease in accuracy. Such results confirm the relevance for site specific samples as indicated before by Vetrivel et al. (2017). Taking advantage of more image data depicting damaged areas coming from other resolutions, instead of independently considering each of the different resolution levels, is beneficial, especially considering the limitations in the quantity of available image data.

Chapters 2 and 3 aimed at using multi-resolution imagery, i.e. from different platforms, to assess its impact on each of the platforms considered (satellite and aerial, both manned and unmanned). The approaches were based on a modified version of the *resnet* (He et al., 2016) architecture and tested on more than 15 different locations, from 4 different continents. The approach relied on a computationally expensive network (*resnet*, with dozens of millions of parameters), which would require dedicated computational power to be of use in an operational context. In the meantime other networks were proposed for image recognition tasks such as the *densenet* (Huang et al., 2017). *Densenet* not only improved the overall accuracy metrics but also the

computational cost.It is likely that in the future different architectures are proposed which further improve both the accuracy and the computational cost of present networks. Hence, a drawback of this approach was to be tailored to a given architecture instead of considering a broader approach which could be more architecture independent. In spite of the use of a multitude of data, given the multi-resolution nature of the approach where damaged samples are considered from a larger set of data (from mono- to multi-resolution), there is often lack of image data to generate reliable damage detection routines. To this regard overwhelming difference of the amount of imagery depicting not damaged areas *vs* damaged ones could be taken advantage of. For example by learning the feature representation of such images depicting non-damaged areas (Oza and Patel, 2019) in a one-class (not damaged) classification approach, which could then be merged with the smaller amount of damaged image samples. This large amount of imagery depicting non-damaged areas could also be used alongside an image classification framework which could successfully address the class imbalance issue (Buda et al., 2018) and take advantage of it. Given that the generated models can only learn from what is present in the training images, an inventory of the different damage evidences present in the training/testing data should be performed. Such inventory should also contain location and image capture details for example. Only then and with an extensive description of a given image dataset, such analysis can be performed. This extensive description of the dataset could then be used to label the damage evidences into different typologies of damage (e.g. partial and total collapse, blown out debris). This could be the input for a multi-class classification or even within a multi-task learning approach (Bittner et al., 2019). This would give an extra information to stakeholders moving from the traditional binary nature of damage detection procedures.

The mapping of debris and rubble piles might leave out smaller damage evidences such as spalling and cracks, especially in the façades. To survey building façades for damage, oblique imagery capturing these building elements is critical, given that otherwise such damage can only be inferred by blown out debris for example. Hence, in this research both aerial manned and UAV have been used to detect façade damages. Using UAV, the focus of the presented research was to make the façade damage detection more efficient due to the usual capture of a larger number of images. Applying a damage detection algorithm to all the images would not be optimal. The objective was to reduce the number of images and image regions to be fed to the damage detection algorithm with the objective of having a faster façade damage detection using UAV. The general idea was to detect the façades and then use that information to extract their respective image patch from the images, while discarding the rest of the image regions. First, the point cloud of tie points was used to detect the building roofs. While many contributions focused on building roof segmentation from point clouds, also considering photogrammetric point clouds and image information (Vetrivel et al., 2015), in this case the objective

was to use the point cloud of tie points instead of the dense image matching point cloud to decrease the computational cost of the approach. This was found to be possible due to the often centimeter resolution of UAV surveys, which was translated in a high concentration of tie points. Nonetheless, points observed at least on three images were considered. This allowed to rely only on stronger tie points, while discarding weaker ones to detect building roofs recurring to point cloud segmentation approaches. With the building roofs detected, the façades were defined and extracted from the oblique images using the raw orientation information coming from the global navigation satellite systems (GNSS) and inertial measurement units (IMU) present onboard the UAV. These façade image patches were then fed to a damage detection procedure trained with debris and rubble piles. For example, with the image dataset used for testing of the approach, only oblique images containing façades were considered and only the façade image patches present in each of the images were fed to a damage detection approach. Given that the damage detection algorithm was mostly trained on nadir images and with image samples of debris and rubble piles, it was prone to a high rate of false positives. It was clear from these experiments that a damage detection procedure tailored for the façades was needed. While more research is needed to build a façade damage detection algorithm; the façade extraction procedure could already be used within an operational context to reduce the amount of images and image regions to be fed to a given façade damage detection algorithm.

This optimization of the façade damage detection approach using the UAV could be combined with a recent framework proposed by Nex et al. (2019). The latter aimed at autonomously and near-real-time mapping of debris and rubble piles using a UAV. Such a system was tailored for FR when surveying a building block for damaged buildings within the INACHUS project. The framework received positive feedback by the FR evaluating two pilot tests performed within INACHUS. The approach specifically aimed at generating an orthophoto (using only nadir images) of the area of interest with the regions presenting debris and rubble piles overlaid in red. However, no attention was given to the façades. The proposed approach in this thesis (chapter 4) could extend the nadir-only mapping of debris and rubble piles to the detection of damaged façades using oblique views, given the focus of the presented approach in the computational cost of such assessment. This would give more information to FR, also regarding the façades still standing in a given area. Moreover, it could be continuously performed to generate constant updates regarding a given building block. All this damage information generated both from nadir and oblique views and focusing on debris and façade damages would certainly enable more informed decisions.

The façade damage detection was then performed using as input aerial oblique imagery captured from manned platforms. These platforms often capture imagery at a lower resolution when comparing with UAV but can cover larger

areas. Moreover, from the previous study, it was clear that an approach trained with debris and rubble piles fell short when applying it for the specific case of the façades, where damage evidences such as cracks, spalling and other smaller signs of damage were overlooked. Given the interest for aerial oblique data, also from damage map producers, we focused on a multi-temporal approach. A preliminary set of tests was performed making use of pre- and post-event aerial oblique imagery. To this end a simple correlation coefficient was proposed to compare pre- and post-event façade image patches. First the comparison was performed intra-epoch (pre-event only), to define a baseline correlation value between different views of a given façade. This was then compared with the correlation coefficient between pre- and post-event façade image patches. On one hand intact façades presented similar correlation coefficients when calculated between pre-event only façade image patches and when calculated between pre- and post-event façade image patches. On the other hand, damaged façades presented a lower correlation coefficient when calculated between pre- and post-event façade image patches. This difference varied regarding each of the façades (e.g. façades with no texture where the correlation coefficient fails), hence to establish a manually defined threshold makes the transferability of the method difficult.

In order to derive a more generalizable approach for the multi-temporal façade damage assessment a supervised classification was tested in chapter 6. The focus of the approach was to derive a framework that given a set of pre- and post-event façade image patches could determine if such façade was intact or damaged. Specifically, the focus of the experiments was two-fold: 1) determine the optimal merge of the multi-temporal imagery within a CNN for façade damage detection, 2) take advantage of the redundancy of aerial manned surveys to extract several façade images patches per façade and embed this with 1). The multi-temporal approaches were compared with mono-temporal approaches which were considered as baseline. The multi-temporal approaches clearly outperformed the mono-temporal ones (up to 25% difference in f1-score). This confirms the general improvement of multi-temporal approaches when compared with mono-temporal ones in remote sensing studies (Tewkesbury et al., 2015). The different views extracted from the multi-temporal image data were combined following two different approaches where image pairs (pre- and post-event façade image patches) and image sextuples (three pre- and three post-event façade image patches) were used as input. With this input several approaches were defined, aimed at better understanding how to consider the different features derived from the façade image patches. Early and late fusion were tested, where it was observed that when using image pairs, it is preferable to perform late fusion and have an independent set of features per epoch. Such experiment was only marginally outperformed by an approach using the image sextuples as input and also merging the different epochs feature maps in a later stage of the network. Hence, per epoch feature information still plays a major role in the

identification of façade damages. Sharing features between different epochs is not optimal where in none of the early/fusion experiments using as input both image pairs and image sextuples outperformed the late fusion ones. However, overall, the results were poor, achieving at the most 82% f1-score. This may be due to the low amount of image data, where only around 90 damaged façades were identified, and in cases where damage evidences were smaller cracks on the wall or other small evidences of damage (i.e. low resolution of the imagery). Moreover, and from an operational point of view, such approach would need to be tested for geographical transferability given that the reported results only refer to a specific Italian region. Aerial manned platforms can survey regional/city wide areas; however, the low resolution of the images makes it challenging to identify smaller signs of damage. To this regard, UAV could be used to survey the most critical and/or occluded façades, capturing image data at a higher resolution and at specific locations, as in chapter 3.

While focusing on the façades, such a framework could be used alongside other studies that focus on the identification of debris and rubble piles, as in chapter 2 and 3. This merge would allow to generate a more complete damage assessment over a given region and, in the aerial case, using the same survey with both nadir and oblique views. It would also enable more certainty in the damage identification since these two sets of results (debris and façade damage) could be compared with one another. Such redundancy is especially relevant for FR, where the damage results need to be reliable. On the other hand, the binary nature of the approach, which includes several typologies of damage is not optimal given the focus of FR on partially and totally collapsed buildings. Like in the identification of rubble piles and debris, there is a need to have not only the localization of the damaged façades but also a qualitative analysis regarding the type of damage that was detected. This would not only increase the information given to stakeholders but would also allow to analyse where such damage detection approaches fail. The experiments performed in this thesis relied heavily on commercial imagery making the public dissemination of such data not possible. This is a bottleneck to researchers in the field, especially when considering meter resolution satellite and/or aerial (manned platforms) imagery. Instead, researchers are disseminating the weights of damage detection networks (Nex et al., 2019). This is not optimal given that there is no raw data to perform experiments but these may be used to, for example, compare different approaches.

Overall, the findings reported in this thesis can be of use in both the response and, recovery and rehabilitation phase of the disaster management cycle. Taking advantage of image data coming from several systems instead of focusing on each of them separately might be beneficial for any actor which aims at mapping debris and rubble piles from remote sensing images. The mapping of façade damages reported in this study could also identify damaged façades from multi-temporal imagery obtained using aerial manned platforms.

However, in both cases (debris and façade damage) there is little information regarding the actual damage evidence that was detected, given the binary nature of the approaches. While the location of the damaged areas/façades is relevant for several actors dealing with disaster management, it falls short in describing the actual damages. Such information is critical, namely in the façade case, given that a collapsed façade entails different actions when comparing with a façade which only contains spalling and/or cracks. Hence, more research is needed to this regard, moving beyond the binary nature of the approaches reported in this study.

7.1 *References*

Bittner, K., Körner, M., Fraundorfer, F., Reinartz, P., 2019. Multi-task cGAN for simultaneous spaceborne DSM refinement and roof-type classification. Remote Sens. 11, 1262. https://doi.org/10.3390/rs11111262

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Dubois, D., Lepage, R., 2014. Fast and efficient evaluation of building damage from very high resolution optical satellite images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 4167–4176. https://doi.org/10.1109/JSTARS.2014.2336236

Fernandez Galarreta, J., Kerle, N., Gerke, M., 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Nat. Hazards Earth Syst. Sci. 15, 1087–1101. https://doi.org/10.5194/nhess-15-1087-2015

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR. https://doi.org/10.1109/CVPR.2016.90

Huang, G., Liu, Z., Maaten, L. van der, Weinberger, K.Q., 2017. Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Nex, F., Duarte, D., Steenbeek, A., Kerle, N., 2019. Towards Real-Time Building Damage Mapping with Low-Cost UAV Solutions. Remote Sens. 11, 287. https://doi.org/10.3390/rs11030287

Nex, F., Duarte, D., Tonolo, F., Kerle, N., 2019. Structural building damage detection with deep learning: assessment of state-of-the-art CNN in operational conditions Remote Sens. 11, 2765. https://doi.org/10.3390/rs11232765

Oza, P., Patel, V.M., 2019. One-Class Convolutional Neural Network. IEEE Signal Process. Lett. 26, 277–281. https://doi.org/10.1109/LSP.2018.2889273

Tewkesbury, A.P., Comber, A.J., Tate, N.J., Lamb, A., Fisher, P.F., 2015. A critical synthesis of remotely sensed optical image change detection techniques. Remote Sens. Environ. 160, 1–14. https://doi.org/10.1016/j.rse.2015.01.006

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2017. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Vetrivel, A., Gerke, M., Kerle, N., Vosselman, G., 2015. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. ISPRS J. Photogramm. Remote Sens. 105, 61–78. https://doi.org/10.1016/j.isprsjprs.2015.03.016

Bibliography

Diogo Duarte was born on 27th December, 1987 in Pombal, Portugal. He received a Bachelors in Civil engineering and a master's diploma in Geomatics from the University of Coimbra (2014), Portugal. The master thesis title was "Automatic orthomosaic production and 3D modelling of urban aeras using point clouds obtained with photogrammetric open source software''. After graduating he enrolled as a research fellow in Coimbra. The research was focused on assessing the photovoltaic protentional of Coimbra roofs for the EMSURE project (2014). He then joined ITC in 2015 as a PhD candidate funded through a FP7 project INACHUS. His research focus on the extraction of information from optical remote sensing images. Worked on several areas such as remote sensing image classification, computer vision, photogrammetry and machine learning.

Author's publications

Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1636.

Duarte, D., Nex, F., Kerle, N., and Vosselman, G.: Towards a more efficient detection of earthquake induced façade damages using oblique UAV imagery, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., **2017**, XLII-2/W6, 93-100,

Duarte, D., Nex, F., Kerle, N., and Vosselman, G.: Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., **2018**, IV-2, 89-96,

Duarte, D., Nex, F., Kerle, N., and Vosselman, G.: Damage detection on building façades using multi-temporal aerial oblique imagery*, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.,* **2019***, IV-2/W5, 29-36*

Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Detection of seismic façade damages with multi-temporal aerial oblique imagery. (under review)

Nex, F.; **Duarte, D**.; Steenbeek, A.; Kerle, N. Towards real-time building damage mapping with low-Cost UAV solutions. *Remote Sens.* **2019**, *11*, 287.

Nex, F.; **Duarte, D**.; Tonolo, F.; Kerle, N. Structural building damage detection with deep learning: assessment of state-of-the-art CNN in operational conditions. *Remote Sens.* **2019**, *11*, 2765.

Vetrivel, A., **Duarte, D**., Nex, F., Gerke, M., Kerle, N., and Vosselman, G.: Potentional of multi-temporal oblique airborne imagery for structural damage assessment, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., **2016**, III-3, 355-362.

Kerle, N., Nex, F., **Duarte, D.**, Vetrivel, A. UAV-based structural damage mapping – results from 6 years of research in two European projects, *Gi4DM,* **2019**